DOCUMENT RESUME

ED 471 955 IR 058 520

TITLE	The State of Digital Preservation: An International Perspective. Conference Proceedings (1st, Washington, D.C., April 24-25, 2002).
INSTITUTION	Council on Library and Information Resources, Washington, DC.
ISBN	ISBN-1-887334-92-0
PUB DATE	2002-07-00
NOTE	103p.; The presentation entitled "How Warner Brothers Is Approaching the Preservation of Its Digital Content" was omitted because Warner Brothers did not approve it for publication. Conference supported by a grant from Documentation Abstracts, Inc.
AVAILABLE FROM	Council on Library and Information Resources, 1755 Massachusetts Ave., NW, Suite 500, Washington, DC 20036 ($20 per copy; orders must be placed through Web site). Tel: 202-939-4750; Fax: 202-939-4765; e-mail: info@clir.org; Web site: http://www.clir.org. For full text: http://www.clir.org/pubs/abstract/pub107abst.html.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	EDRS Price MF01/PC05 Plus Postage.
DESCRIPTORS	*Archives; Conference Proceedings; Electronic Libraries; Foreign Countries; Library Development; Library Technical Processes; *Preservation
IDENTIFIERS	Australia; Digital Data; *Digital Preservation; *Digital Technology; Netherlands; United States

ABSTRACT

In this collection of papers presented at "The State of Digital Preservation: An International Perspective" conference (Washington, DC, April 24-25, 2002), leading experts from the United States, the Netherlands, and Australia describe current practices and challenges in digital preservation. Contents include: "Introduction: The Changing Preservation Landscape" (Deanna Marcum); "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years" (Kenneth Thibodeau); "The Digital Preservation Research Agenda" (Margaret Hedstrom); "Understanding Digital Preservation: A Report from OCLC" (Meg Bellinger); "Update on the National Digital Infrastructure Initiative" (Laura Campbell); "Experience of the National Library of the Netherlands" (Titia van der Werf); "Digital Preservation--A Many-Layered Thing: Experience at the National Library of Australia" (Colin Webb); and "Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information" (Donald Waters). (AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 471 955

The State of Digital Preservation: An International Perspective

CONFERENCE PROCEEDINGS
DOCUMENTATION ABSTRACTS, INC.
INSTITUTES FOR INFORMATION SCIENCE
WASHINGTON, D.C.
APRIL 24-25, 2002

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B.H. Leney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR058520

COUNCIL ON LIBRARY AND INFORMATION RESOURCES

July 2002

BEST COPY AVAILABLE

The State of Digital Preservation: An International Perspective

CONFERENCE PROCEEDINGS
DOCUMENTATION ABSTRACTS, INC.
INSTITUTES FOR INFORMATION SCIENCE
WASHINGTON, D.C.
APRIL 24-25, 2002

Council on Library and Information Resources
Washington, D.C.

July 2002

Documentation Abstracts, Inc. Institutes for Information Science

"The State of Digital Preservation: An International Perspective" is the first in a series of international symposiums that are supported by a grant from Documentation Abstracts, Inc. (DAI). The institutes, presented by the Council on Library and Information Resources (CLIR) will address key issues in information science relating to digital libraries, economics of information, or resources for scholarship.

Documentation Abstracts, Inc. was established in 1966 as a nonprofit organization comprising representatives from eight societies in the field of library and information science: American Chemical Society–Division of Chemical Information, American Library Association, American Society of Indexers, American Society for Information Science and Technology, Association of Information and Dissemination Centers, Association for Library and Information Science Education, Medical Library Association, and Special Libraries Association.

DAI was established to organize, evaluate, and disseminate information and knowledge concerning the various aspects of information science. It did this through publishing *Information Science Abstracts (ISA)*, a bimonthly abstracting and indexing publication covering the literature of information science worldwide. In June 1998, this periodical was acquired by Information Today, Inc., which continues its publication to date.

The Council on Library and Information Resources is an independent, nonprofit organization dedicated to improving the management of information for research, teaching, and learning. CLIR works to expand access to information, however recorded and preserved, as a public good.

CLIR's agenda is framed by a single question: What is a library in the digital age? Rapid changes in technology, evolving intellectual property legislation, new modes of scholarly communication, and new economic models for information provision have all contributed to a new information environment for libraries. In partnership with other organizations, CLIR helps create services that expand the concept of "library" and supports the providers and preservers of information.

ISBN 1-887334-92-0

Published by:

Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036
Web site at http://www.clir.org

Additional copies are available for $20 per copy. Orders must be placed through CLIR's Web site.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Copyright 2002 by the Council on Library and Information Resources. No part of this publication may be reproduced or transcribed in any form without permission of the publisher. Requests for reproduction should be submitted to the Director of Communications at the Council on Library and Information Resources.

Contents

About the Authors iv

Introduction: The Changing Preservation Landscape, *Deanna Marcum* 1

Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, *Kenneth Thibodeau* 4

Work Being Done to Understand Digital Preservation: Project Reports

The Digital Preservation Research Agenda, *Margaret Hedstrom* 32

Understanding Digital Preservation: A Report from OCLC, *Meg Bellinger* 38

Update on the National Digital Infrastructure Initiative, *Laura Campbell* 49

International Initiatives

Experience of the National Library of the Netherlands, *Titia van der Werf* 54

Digital Preservation—A Many-Layered Thing: Experience at the National Library of Australia, *Colin Webb* 65

Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information, *Donald Waters* 78

CLIR regrets that this volume of proceedings omits the presentation entitled "How Warner Brothers is Approaching the Preservation of its Digital Content," which Warner Brothers did not approve for publication.

About the Authors

Meg Bellinger joined Preservation Resources, then an OCLC subsidiary, as president in 1993. She was promoted to vice president of OCLC Digital and Preservation Resources in 2001. Ms. Bellinger came to OCLC from Research Publications International in Woodbridge, Connecticut, where she held management positions in the company's product development and editorial departments before being promoted to vice president of editorial development and preservation in 1991. She is a member of the board of the UK Digital Preservation Coalition and of the Digital Library Federation Steering Committee.

Laura Campbell is associate librarian for strategic initiatives at the Library of Congress (LC). She is responsible for strategic planning for LC, a task that currently includes the development of a national strategy, in cooperation with other institutions, for the collection, access, and preservation of digital materials. Ms. Campbell also oversees LC's Information Technology Services directorate. Before joining LC in 1992, she worked several years in the private sector, serving as vice president at QueTel Corp., and as manager and principal for Arthur Young & Co. (currently Ernst & Young).

Margaret Hedstrom is associate professor at the School of Information, University of Michigan, where she teaches in the areas of archives, electronic records management, and digital preservation. Her chief research interests at present are record keeping in collaborative work environments and methods for long-term preservation of complex digital objects. She is project director for the CAMiLEON Project, an international research project to investigate the feasibility of emulation as a digital preservation strategy. She was a member of the National Research Council study committee that prepared *LC21*, a report on the digital future of the Library of Congress. She is a member of its National Digital Strategy Advisory Board.

Deanna Marcum is president of the Council on Library and Information Resources (CLIR), formed in 1997 by the merger of the Commission on Preservation and Access and the Council on Library Resources. From 1995 to 1997, she served as president of both organizations simultaneously. Her career has included tenure as director of public service and collection management at the Library of Congress and dean of the School of Library and Information Service at The Catholic University of America. From 1980 to 1989, she was first a program officer and then vice president of the Council on Library Resources.

Kenneth Thibodeau is director of the Electronic Records Archives Program at the National Archives and Records Administration (NARA). He has more than 25 years' experience in archives and records management and is an internationally recognized expert in electronic records. He organized and directed the Center for Electronic Records at the NARA from 1988 to 1998. In 1996, he was detailed from NARA to serve as the director of the Department of Defense Records Management Task Force, a group established to implement business processing reengineering of records management.

Donald Waters is the program officer for scholarly communications at The Andrew W. Mellon Foundation. Before joining the Foundation, he served as the first director of the Digital Library Federation, as associate university librarian at Yale University, and in a variety of other positions at the Computer Center, the School of Management, and the University Library at Yale. In 1995-1996, he cochaired the Task Force on Archiving of Digital Information, and was the editor and a principal author of the Task Force Report. He is the author of numerous articles and presentations on libraries, digital libraries, digital preservation, and scholarly communications.

Titia van der Werf was project manager at the Library Research and Development Department of the Koninklijke Bibliotheek, National Library of the Netherlands, from 1993 to 2002. She initiated new Internet-based library products and services that have grown into full operational services, such as the library Web site, the Dutch academic subject gateway DutchESS, and the Web-guide NL-menu. Since 1996, she was involved primarily in digital archiving and long-term preservation. She coordinated the NEDLIB project and contributed to the implementation of the library's deposit system with IBM Netherlands. In July 2002, she left the National Library of the Netherlands to become head of the library at the African Studies Centre in Leiden.

Colin Webb is director of preservation services at the National Library of Australia, where he has worked since 1993. After professional training as a bookbinder and as a book, paper, and photographic conservator, he worked for the National Archives of Australia as a preservation manager for more than a decade before moving to the National Library to set up a program in information preservation. He created the first—and still the only—specialist positions in digital preservation in Australia, and he has sought to bring a strong preservation perspective to the National Library's digital initiatives.

Introduction: The Changing Preservation Landscape

Deanna Marcum

The Council on Library and Information Resources (CLIR) and, later, the Digital Library Federation (DLF) have been exploring the topic of preserving digital information for a long time. Don Waters and John Garrett wrote their landmark report, *The Preservation of Digital Information*, in 1996. In describing the problem, they wrote

> Rapid changes in the means of recording information, in the formats for storage, and in the technologies for use threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, "nasty, brutish, and short."
>
> Today, information technologies that are increasingly powerful and easy to use, especially those that support the World Wide Web, have unleashed the production and distribution of digital information. . . . If we are effectively to preserve for future generations the portion of this rapidly expanding corpus of information in digital form that represents our cultural record, we need to understand the costs of doing so and we need to commit ourselves technically, legally, economically, and organizationally to the full dimensions of the task. Failure to look for trusted means and methods of digital preservation will certainly exact a stiff, long-term cultural penalty.

In the summary of their report, Waters and Garrett concluded that

- The first line of defense against loss of valuable digital information rests with the creators, producers, and owners of that information.

- A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections.
- A process of certification for digital archives is needed to create a climate of trust about the prospects of preserving digital information.
- Certified digital archives must have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism for preserving valuable digital information that is in jeopardy of destruction, neglect, or abandonment by its current custodian.

These conclusions were reached after an 18-month study by a task force composed of librarians, archivists, technologists, government officials, publishers, creators, lawyers, and museum directors. The group issued nine recommendations in the areas of pilot projects, needed support structures, and best practices.

Six years later, what is the state of preservation of digital information? We have looked at many institutions and organizations to understand what has been accomplished.

Our first observation is that a great variety of projects have been undertaken, both in the United States and in other parts of the world. I cannot begin to describe all that is being done, but will list some significant work that has been done since 1996.

- In the United Kingdom, a Digital Preservation Coalition has been established.
- The National Library of Australia has established PADI (Preserving Access to Digital Information), a subject gateway to digital preservation resources.
- CLIR and the DLF have published several reports designed to increase awareness of the problem and what research is being done to address it.
- Organizations have worked hard to establish standards and best practices. The Online Computer Library Center (OCLC) and the Research Libraries Group jointly have developed two working documents to establish best practices: *Attributes of a Digital Archive for Research Repositories* and *Preservation Metadata for Long-Term Retention.*
- Practical experiments have been funded. The Andrew W. Mellon Foundation has funded seven universities to work with publishers to plan for digital repositories for e-journal content. Through PubMed Central, the National Library of Medicine acts as a digital archival repository for medical publications and other medical information.
- The Library of Congress is developing a national strategy for preserving digital information. With an extra appropriation of $100 million from the U.S. Congress, the Library has formed a national advisory board and is working with a number of governmental and private agencies to develop this plan.

- The commercial and entertainment sectors have made great advances in understanding digital preservation, because they must manage their digital assets if they are to have products in the future.

Our aim in organizing this first DAI Institute for Information Science was to look at some of the most interesting developments in the preservation of digital information. We hoped that by bringing together so much talent, we could identify some of the barriers that impede progress and figure out ways to overcome them. The symposium speakers provided a rich mix of lessons learned, perspectives on recent developments, and analysis of the challenges ahead. These are reflected in the following pages. We are grateful to each presenter for helping advance the discussion and leaving us with much food for thought.

We are also deeply grateful to Documentation Abstracts, Inc. (DAI), which has provided support for CLIR to organize a new series of symposia on timely information science topics. We are encouraged by the success of this first program and look forward to subsequent symposiums in the DAI Information Institute series.

Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years[1]

Kenneth Thibodeau

What Does It Mean to Preserve Digital Objects?

The preservation of digital objects involves a variety of challenges, including policy questions, institutional roles and relationships, legal issues, intellectual property rights, and metadata. But behind or perhaps beneath such issues, there are substantial challenges at the empirical level. What does it mean to preserve digital objects? What purposes are served by preserving them? What are the real possibilities for successful preservation? What are the problems encountered in trying to exploit these possibilities? Can we articulate a framework or an overall architecture for digital preservation that allows us to discriminate and select possibilities?

To address any of these challenges, we must first answer the simple question: What are digital objects? We could try to answer this question by examining the types of digital objects that have been and are being created. Many types of digital information can and do exist in other forms. In fact, many types of digital information are rather straightforward transcriptions of traditional documents, such as books, reports, correspondence, and lists. Other types of digital information are variations of traditional forms. But many forms of digital information cannot be expressed in traditional hard-copy or analog media; for example, interactive Web pages, geographic information systems, and virtual reality models. One benefit of an extensive review of the variety of types of digital information is that it forces one to come to grips with this variety, which is growing both in terms of the number of types of digital objects and in terms of their complexity.

---

[1] An earlier version of this paper appeared as "Digital Preservation Techniques: Evaluating the Options" in *Archivi & Computer: Automatione e Beni Culturali* 10 (2/01): 101–109.

In fact, the diversity of digital information exists not only among types but also within types. Consider one application class, documents. There is no single definition or model of a digital document that would be valid in all cases. Information technologists model digital documents in very different ways: a digital document can be a sequence of expressions in natural language characters or a sequence of scanned page images, a directed graph whose nodes are pages, what appears in a Web page, and so on. How documents are managed, and therefore how they are preserved, depend on the model that is applied.

The variety and complexity of digital information objects engender a basic criterion for evaluating possible digital preservation methods, namely, they must address this variety and complexity. Does that necessarily mean that we must preserve the variety and complexity? It is tempting to respond that the variety and complexity must indeed be preserved because if we change the characteristics of digital objects we are obviously not preserving them. However, that response is simplistic. For example, in support of the argument that emulation is the best method for digital preservation—because it allows us to keep digital objects in their original digital formats—the example of the periodic table of the elements has been offered. The information conveyed by the periodic table depends on the spatial layout of the data contained in it. The layout can be corrupted or obliterated by using the wrong software, or even by changing the font. However, to argue that any software or digital format is necessary to preserve the periodic table is patently absurd. The periodic table was created a century before computers, and it has survived very well in analog form. Thus we cannot say without qualification that the variety and complexity of digital objects must always be preserved. In cases such as that of the periodic table, it is the essential character of the information object, not the way it happens to be encoded digitally, that must be preserved. For objects such as the periodic table, one essential characteristic is the arrangement of the content in a 2-by-2 grid. As long as we preserve that structure, we can use a variety of digital fonts and type sizes, or no fonts at all—as in the case of ASCII or a page-image format.

We can generalize this insight and assert that the preservation of a digital information object does not necessarily entail maintaining all of its digital attributes. In fact, it is common to change digital attributes substantially to ensure that the essential attributes of an information object are preserved when the object is transmitted to different platforms. For example, to ensure that written documents retain their original appearance, authors translate them from the word processing format in which they were created to Adobe's PDF format. Fundamentally, the transmission of information objects across technological boundaries—such as platforms, operating systems, and applications—is the same, whether the boundaries exist in space or time.

Are there basic or generic properties that are true of all digital objects? From a survey of types such as those just described, one

could derive an intensive definition of digital objects: a digital object is an information object, of any type of information or any format, that is expressed in digital form. That definition may appear too generic to be of any use in addressing the challenge of digital preservation. But if we examine what it means for information to be expressed in digital form, we quickly come to recognize a basic characteristic of digital objects that has important consequences for their preservation. All digital objects are entities with multiple inheritance; that is, the properties of any digital object are inherited from three classes. Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different. A *physical* object is simply an inscription of signs on some physical medium. A *logical* object is an object that is recognized and processed by software. The *conceptual* object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions.

Physical Objects: Signs Inscribed on a Medium

As a physical object, a digital object is simply an inscription of signs on a medium. Conventions define the interface between a system of signs, that is, a way of representing data, and the physical medium suitable for storing binary inscriptions. Those conventions vary with the physical medium: there are obvious physical differences between recording on magnetic disks and on optical disks. The conventions for recording digital data also vary within media types; for example, data can be recorded on magnetic tape with different densities, different block sizes, and a different orientation with respect to the length and width of the tape.

Basically, the physical level deals with physical files that are identified and managed by some storage system. The physical inscription is independent of the meaning of the inscribed bits. At the level of physical storage, the computer system does not know what the bits mean, that is, whether they comprise a natural language document, a photograph, or anything else. Physical inscription does not entail morphology, syntax, or semantics.

Concern for physical preservation often focuses on the fact that digital media are not durable over long periods of time (Task Force 1996). This problem can be addressed through copying digital information to new media, but that "solution" entails another type of problem: media refreshment or migration adds to the cost of digital preservation. However, this additional cost element may in fact reduce total costs. Thanks to the continuing operation of Moore's law, digital storage densities increase while costs decrease. So, repeated copying of digital data to new media over time reduces per-unit costs. Historically, storage densities have doubled and costs decreased by half on a scale of approximately two years. At this rate, media migration can yield a net reduction, not an increase, in operational costs: twice the volume of data can be stored for half the cost

(Moore et al. 2000). In this context, the durability of the medium is only one variable in the cost equation: the medium needs to be reliable only for the length of time that it is economically advantageous to keep the data on it. For example, if the medium is reliable for only three years, but storage costs can be reduced by 50 percent at the end of two years, then the medium is sufficiently durable in a preservation strategy that takes advantage of the decreasing costs by replacing media after two years.

The physical preservation strategy must also include a reliable method for maintaining data integrity in storage and in any change to storage, including any updating of the storage system, moving data from inactive storage to a server or from a server to a client system, or delivering information to a customer over the Internet, as well as in any media migration or media refreshment.

Obviously, we have to preserve digital objects as physical inscriptions, but that is insufficient.

Logical Objects: Processable Units

A digital information object is a logical object according to the logic of some application software. The rules that govern the logical object are independent of how the data are written on a physical medium. Whereas, at the storage level, the bits are insignificant (i.e., their interpretation is not defined), at the logical level the grammar is independent of physical inscription. Once data are read into memory, the type of medium and the way the data were inscribed on the medium are of no consequence. The rules that apply at the logical level determine how information is encoded in bits and how different encodings are translated to other formats; notably, how the input stream is transformed into the system's memory and output for presentation.

A logical object is a unit recognized by some application software. This recognition is typically based on data type. A set of rules for digitally representing information defines a data type. A data type can be primitive, such as ASCII or integer numbers, or it can be composite—that is, a data type composed of other data types that themselves might be composite. The so-called "native formats" produced by desktop application software are composite data types that include ASCII and special codes related to the type of information objects the software produces; for example, font, indentation, and style codes for word processing files. A string of data that all conform to the same data type is a logical object. However, the converse is not necessarily true: logical objects may be composite, i.e., they may contain other logical objects.

The logical string must be stored in a physical object. It may be congruent with a physical object—for example, a word processing document may be stored as a single physical file that contains nothing but that document—but this is not necessarily the case. Composite logical objects are an obvious exception, but there are other exceptions as well. A large word processing document can be divided into

subdocuments, with each subdocument, and another object that defines how the subdocuments should be combined, stored as separate physical files. For storage efficiency, many logical objects may be combined in a large physical file, such as a UNIX TAR file. Furthermore, the mapping of logical to physical objects can be changed with no significance at the logical level. Logical objects that had been stored as units within a composite logical object can be extracted and stored separately as distinct physical files, with only a link to those files remaining in the composite object. The way they are stored is irrelevant at the logical level, as long as the contained objects are in the appropriate places when the information is output. This requires that every logical object have its own persistent identifier, and that the location or locations where each object is stored be specified. More important, to preserve digital information as logical objects, we have to know the requirements for correct processing of each object's data type and what software can perform correct processing.

Conceptual Objects: What We Deal with in the Real World

The conceptual object is the object we deal with in the real world: it is an entity we would recognize as a meaningful unit of information, such as a book, a contract, a map, or a photograph. In the digital realm, a conceptual object may also be one recognized by a business application, that is, a computer application that executes business transactions. For example, when you withdraw money from an ATM machine, you conceive of the transaction as an event that puts money in your hands and simultaneously reduces the balance of your bank account by an equal amount. For this transaction to occur, the bank's system that tracks your account also needs to recognize the withdrawal, because there is no human involved at that end. We could say that in such cases the business application is the surrogate or agent for the persons involved in the business transaction.

The properties of conceptual objects are those that are significant in the real world. A cash withdrawal has an account, an account owner, an amount, a date, and a bank. A report has an author, a title, an intended audience, and a defined subject and scope. A contract has provisions, contracting parties, and an effective date. The content and structure of a conceptual object must be contained somehow in the logical object or objects that represent that object in digital form. However, the same conceptual content can be represented in very different digital encodings, and the conceptual structure may differ substantially from the structure of the logical object. The content of a document, for example, may be encoded digitally as a page image or in a character-oriented word processing document. The conceptual structure of a report—e.g., title, author, date, introduction—may be reflected only in digital codes indicating differences in presentation features such as type size or underscoring, or they could be matched by markup tags that correspond to each of these elements. The term "unstructured data" is often used to characterize digital objects that

do not contain defined structural codes or marks or that have structural indicators that do not correspond to the structure of the conceptual object.

Consider this paper. What you see is the conceptual object. Then consider the two images below. Each displays the hexadecimal values of the bytes that encode the beginning of the document.[2] Neither looks like the conceptual object (the "real" document). Neither is the exact equivalent of the conceptual document. Both contain the title of

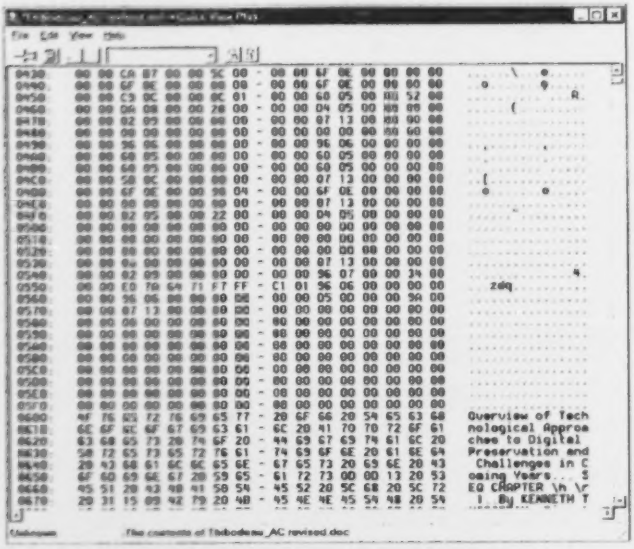


Fig. 1. Hexadecimal Dump of MS Word

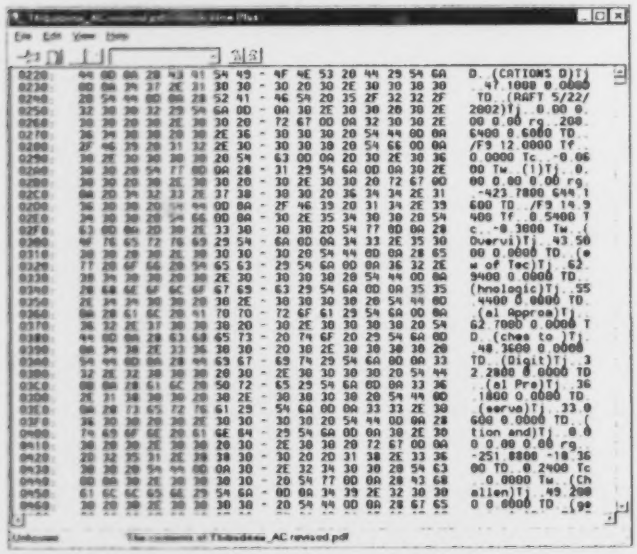


Fig. 2. Hexadecimal Dump of PDF

[2] Each image displays the hexadecimal values of (1) in the leftmost column, the position of first byte in that row relative to the start of the file, and (2) the numeric values of 16 bytes starting with the numbered one. It also shows the printable ASCII characters, or a '.' for unprintable bytes in the rightmost column.

BEST COPY AVAILABLE

the article, but otherwise they differ substantially. Thus, they are two different logical representations of the same conceptual object.

Is there any sense in which we could say that one of these digital formats is the true or correct logical representation of the document? An objective test would be whether the digital format preserves the document exactly as created. The most basic criterion is whether the document that is produced when the digital file is processed by the right software is identical to the original. In fact, each of these encodings, when processed by software that recognizes its data type, will display or print the document in the format in which it was created. So if the requirement is to maintain the content, structure, and visual appearance of the original document, either digital format is suitable. The two images are of Microsoft Word and Adobe PDF versions of the document. Other variants, such as WordPerfect, HTML, and even a scanned image of the printed document, would also satisfy the test of outputting the correct content in the original format.

This example reveals two important aspects of digital objects, each of which has significant implications for their preservation. The first is that there can be different digital encodings of the same conceptual object and that different encodings can preserve the essential characteristics of the conceptual object. The second relates to the basic concept of digital preservation.

With respect to the first of these implications, the possibility of encoding the same conceptual object in a variety of digital formats that are equally suitable for preserving the conceptual object can be extended to more complex types of objects and even to cases where the conceptual object is not presented to a human but is found only at the interface of two business applications. Consider the example of the cash withdrawal from an ATM. The essential record of that transaction consists of information identifying the account from which the cash is withdrawn, the amount withdrawn, and the date and time of the transaction. For the transaction to be carried out, there must be an interface between the system that manages the ATM and the system that manages the account. The information about the transaction presented at the interface, in the format specified for that interface, is the conceptual object that corresponds to the withdrawal slip that would have been used to record the transaction between the account holder and a human teller. The two systems must share that interface object and, in any subsequent actions related to that withdrawal, must present the same information; however, there is no need for the two systems to use identical databases to store the information.

Before considering the implications for the nature of digital preservation, we should examine more fully the relationships among physical, logical, and conceptual objects.

Relationships: Where Things Get Interesting

The complex nature of a digital object having distinct physical, logical, and conceptual properties gives rise to some interesting considerations for digital preservation, especially in the relationships among the properties of any object at these three levels. The relationship between any two levels can be simple. It can be one-to-one; for example, a textual document saved as a Windows word processing file is a single object at all three levels. But a long textual report could be broken down into a master and three subdocuments in word processing format, leaving one conceptual object stored as four logical objects: a one-to-many relationship. If the word processing files relied on external font libraries, additional digital objects would be needed to reproduce the document. Initially, the master and subdocuments would probably be stored in as many physical files, but they might also be combined into a zip file or a Java ARchive (JAR) file. In this case, the relationship between conceptual and logical objects is one-to-many, and the relationship between logical and physical could be either one-to-one or many-to-one. To access the report, it would be necessary to recombine the master and subdocuments, but this amalgamation might occur only during processing and not affect the retention of the logical or physical objects.

Relationships may even be many-to-many. This often occurs in databases where the data supporting an application are commonly stored in multiple tables. Any form, report, or stored view defined in the application is a logical object that defines the content, structure, and perhaps the appearance of a class of conceptual objects, such as an order form or a monthly report. Each instance of such a conceptual object consists of a specific subset of data drawn from different tables, rows, and columns in the database, with the tables and columns specified by the form or report and the rows determined in the first instance by the case, entity, event, or other scope specified at the conceptual level, e.g., order number, "x"; or monthly report for customer, "y"; or product, "z." In any instance, such as a given order, there is a one-to-many relationship between the conceptual and the logical levels, but the same set of logical objects (order form specification, tables) is used in every instance of an order, so the relationship between conceptual and logical objects are in fact many-to-many. In cases such as databases and geographic information systems, such relationships are based on the database model, but many-to-many relationships can also be established on an ad hoc basis, such as through hyperlinks to a set of Web pages or attachments to e-mail messages. Many-to-many relationships can also exist between logical and physical levels; for example, many e-mail messages may be stored in a single file, but attachments to messages might be stored in other files.

To preserve a digital object, the relationships between levels must be known or knowable. To retrieve a report stored as a master and several subdocuments, we must know that it is stored in this fashion and we must know the identities of all the logical components. To retrieve a specific order from a sales application, we do not

need to know where all or any of the data for that order are stored in the database; we only need to know how to locate the relevant data, given the logical structure of the database.

We can generalize from these observations to state that, in order to preserve a digital object, we must be able to identify and retrieve all its digital components. The digital components of an object are the logical and physical objects that are necessary to reconstitute the conceptual object. These components are not necessarily limited to the objects that contain the contents of a document. Digital components may contain data necessary for the structure or presentation of the conceptual object. For example, font libraries for character-based documents and style sheets for HTML pages are necessary to preserve the appearance of the document. Report and form specifications in a database application are necessary to structure the content of documents.

In addition to identifying and retrieving the digital components, it is necessary to process them correctly. To access any digital document, stored bit sequences must be interpreted as logical objects and presented as conceptual objects. So digital preservation is not a simple process of preserving physical objects but one of preserving the ability to reproduce the objects. The process of digital preservation, then, is inseparable from accessing the object. You cannot prove that you have preserved the object until you have re-created it in some form that is appropriate for human use or for computer system applications.

To preserve a digital object, is it necessary to preserve its physical and logical components and their interrelationship, without any alteration? The answer, perhaps surprisingly, is no. It is possible to change the way a conceptual object is encoded in one or more logical objects and stored in one or more physical objects without having any negative impact on its preservation. For example, a textual report may contain a digital photograph. The photograph may have been captured initially as a JPEG file and included in the report only by means of a link inserted in the word processing file, pointing to the image file. However, the JPEG file could be embedded in the word processing file without altering the report as such. We have seen another example of this in the different formats that can be used to store and reproduce this article. In fact, it may be beneficial or even necessary to change logical or physical characteristics to preserve an object. Authors often transform documents that they create as word processing documents into PDF format to increase the likelihood that the documents will retain their original appearance and to prevent users from altering their contents. An even simpler case is that of media migration. Digital media become obsolete. Physical files must be migrated to new media; if not, they will become inaccessible and will eventually suffer from the physical deterioration of the older media. Migration changes the way the data are physically inscribed, and it may improve preservation because, for example, error detection and correction methods for physical inscription on digital media have improved over time.

Normally, we would say that changing something directly conflicts with preserving it. The possibility of preserving a digital object while changing its logical encoding or physical inscription appears paradoxical and is compounded by the fact that it may be beneficial or even necessary to make such changes. How can we determine what changes are permissible and what changes are most beneficial or necessary for preservation? Technology creates the possibilities for change, but it cannot determine what changes are permissible, beneficial, necessary, or harmful. To make such determinations, we have to consider the purpose of preservation.

The Ultimate Outcome: Authentic Preserved Documents

What is the goal of digital preservation? For archives, libraries, data centers, or any other organizations that need to preserve information objects over time, the ultimate outcome of the preservation process should be authentic preserved objects; that is, the outputs of a preservation process ought to be identical, in all essential respects, to what went into that process. The emphasis has to be on the identity, but the qualifier of "all essential respects" is important.

The ideal preservation system would be a neutral communications channel for transmitting information to the future. This channel should not corrupt or change the messages transmitted in any way. You could conceive of a digital preservation system as a black box into which you can put bit streams and from which you can withdraw them at any time in the future. If the system is trustworthy, any document or other digital object preserved in and retrieved from the system will be authentic. In abstract terms, we would like to be able to assert that, if X_{t0} was an object put into the box at time, t_0, and X_{tn} is the same object retrieved from the box at a later time, tn, then $X_{tn} = X_{t0}$.

However, the analysis of the previous sections shows that this cannot be the case for digital objects. The process of preserving digital objects is fundamentally different from that of preserving physical objects such as traditional books or documents on paper. To access any digital object, we have to retrieve the stored data, reconstituting, if necessary, the logical components by extracting or combining the bit strings from physical files, reestablishing any relationships among logical components, interpreting any syntactic or presentation marks or codes, and outputting the object in a form appropriate for use by a person or a business application. Thus, it is impossible to preserve a digital document as a physical object. One can only preserve the ability to reproduce the document. Whatever exists in digital storage is not in the form that makes sense to a person or to a business application. The preservation of an information object in digital form is complete only when the object is successfully output. The real object is not so much retrieved as it is reproduced by processing the physical and logical components using software that recognizes and properly handles the files and data types (InterPARES Preservation Task Force 2001). So, the black box for digital preserva-

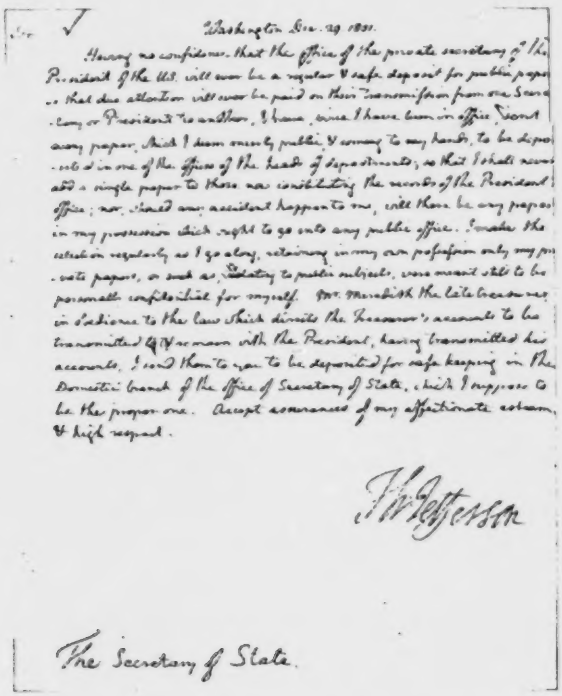
tion is not just a storage container: it includes a process for ingesting objects into storage and a process for retrieving them from storage and delivering them to customers. These processes, for digital objects, inevitably involve transformations; therefore, the equation, then $X_{tn} = X_{t0}$ cannot be true for digital objects.

In fact, it can be argued that practically, this equation is never absolutely true, even in the preservation of physical objects. Paper degrades, ink fades; even the Rosetta Stone is broken. Moreover, in most cases we are not able to assert with complete assurance that no substitution or alteration of the object has occurred over time. As Clifford Lynch has cogently argued, authentication of preserved objects is ultimately a matter of trust. There are ways to reduce the risk entailed by trusting someone, but ultimately, you need to trust some person, some organization, or some system or method that exercises control over the transmission of information over space, time, or technological boundaries. Even in the case of highly durable physical objects such as clay tablets, you have to trust that nobody substituted forgeries over time (Lynch 2000). So the equation for preservation needs to be reformulated as $X_{tn} = X_{t0} + \Delta(X)$, where $\Delta(X)$ is the net effect of changes in X over time.

But can an object change and still remain authentic? Common sense suggests that something either is or is not authentic, but authenticity is not absolute. Jeff Rothenberg has argued that authenticity depends on use (Rothenberg 2000). More precisely, the criteria for authenticity depend on the intended use of the object. You can only say something is authentic with respect to some standard or criterion or model for what X is.

Consider the simple example shown in figure 3. It shows a letter, preserved in the National Archives, concerning the disposition of Thomas Jefferson's papers as President of the United States (Jefferson 1801). Is this an authentic copy of Thomas Jefferson's writing? To answer that question, we would compare it to other known cases of Thomas Jefferson's handwriting. The criteria for authentication would relate to the visual appearance of the text. But what if, by "Jefferson's writing," we do not mean his handwriting but his thoughts? In that case, the handwriting becomes irrelevant: Jefferson's secretary may have written the document, or it could even be a printed version. Conversely, a document known to be in Jefferson's handwriting, but containing text he copied from a book, does not reveal his thoughts. Authenticating Jefferson's writing in this sense relates to the content and style, not to the appearance of the text. So authenticating something as Jefferson's writing depends on how we define that concept.

There are contexts in which the intended use of preserved information objects is well-known. For example, many corporations preserve records for very long times for the purpose of protecting their property rights. In such cases, the model or standard that governs the preservation process is that of a record that will withstand attacks on its reliability and authenticity in litigation. Institutions such as libraries and public archives, however, usually cannot prescribe or predict the uses that will be made of their holdings. Such institutions



Washington Dec. 29 1801.

Having no confidence that the office of the private secretary of the President of the U.S. will ever be a regular & safe depositary for public papers, that due attention will ever be paid to their transmission from one Secretary or President to another, I have, since I have been in office, sorted every paper which I deem merely public & coming to my hands, to be deposited in one of the offices of the heads of departments, so that I shall receive & add a single paper to those now constituting the records of the President's office; nor, should any accident happen to me, will there be any papers in my possession which ought to go into any public office. I make the return regularly as I go along, retaining in my own possession only my private papers, or such as, relating to public subjects, were meant still to be personally confidential for myself. For Murdoch the late treasurer, in obedience to the law which directs the Treasurer's accounts to be transmitted & I presume with the President, having transmitted his accounts, I send them to you to be deposited for safe keeping in the Domestic branch of the office of Secretary of State, which I suppose to be the proper one. Accept assurances of my affectionate esteem, & high respect.

Th: Jefferson

The Secretary of State.

Fig. 3. Jefferson note

generally maintain their collections for access by anyone, for whatever reason. Where the intentions of users are not known in advance, one must take an "aboriginal" approach to authenticity; that is, one must assume that any valid intended use must be somehow consonant with the original nature and use of the object. Nonetheless, given that a digital information object is not something that is preserved as an inscription on a physical medium, but something that can only be constructed—or reconstructed—by using software to process stored inscriptions, it is necessary to have an explicit model or standard that is independent of the stored object and that provides a criterion, or at least a benchmark, for assessing the authenticity of the reconstructed object.

Ways to Go: Selecting Methods

What are the possibilities for preserving authentic digital information objects? Among these possibilities, how can we select the best option or options? Four criteria apply in all cases: any method chosen for preservation must be feasible, sustainable, practicable, and appropriate. *Feasibility* requires hardware and software capable of implementing the method. *Sustainability* means either that the method can be applied indefinitely into the future or that there are credible grounds for asserting that another path will offer a logical sequel to the method, should it cease being sustainable. The sustainability of any given method has internal and external components: internally, the method must be immune or isolated from the effects of technological obsolescence; externally, it must be capable of interfacing

BEST COPY AVAILABLE

with other methods, such as for discovery and delivery, which will continue to change. *Practicality* requires that implementation be within reasonable limits of difficulty and expense. *Appropriateness* depends on the types of objects to be preserved and on the specific objectives of preservation. With respect to the types of objects to be preserved, we can define a spectrum of possibilities running from preserving technology itself to preserving objects that were produced using information technology (IT). Methods can be aligned across this spectrum because the appropriateness of any preservation method depends on the specific objectives for preservation in any given case. As discussed earlier, the purposes served by preservation can vary widely. Considering where different methods fall across this spectrum will provide a basis for evaluating their appropriateness for any given purpose.

To show the rationale of the spectrum, consider examples at each end. On the "preserve technology" end, one would place preserving artifacts of technology, such as computer games. Games are meant to be played. To play a computer game entails keeping the program that is needed to play the game operational or substituting an equivalent program, for example, through reverse engineering, if the original becomes obsolete. On the "preserve objects" end, one would place preserving digital photographs. What is most important is that a photograph present the same image 50 or 100 years from now as it does today. It does not really matter what happens to the bits in the background if the same image can be retrieved reliably. Conversely, if a digital photograph is stored in a physical file and that file is maintained perfectly intact, but it becomes impossible to output the original image in the future—for example, because a compression algorithm used to create the file was either lossy or lost—we would not say the photograph was preserved satisfactorily.

But these illustrations are not completely valid. Many computer games have no parallels in the analog world. Clearly they must be preserved as artifacts of IT. But there are many games now played on computers that existed long before computers were invented. The card game, solitaire, is one example. Obviously, it could be preserved without any computer. In fact, the most assured method for preserving solitaire probably would be simply to preserve the rules of the game, including the rules that define a deck of cards. So the most appropriate method for preserving a game depends on whether we consider it to be essentially an instance of a particular technology—where "game" is inseparable from "computer"—or a form of play according to specified rules; that is, a member of a class of objects whose essential characteristics are independent of the technology used to produce or implement them. We have to preserve a computer game in digital form only if there is some essential aspect of the digital form than cannot be materialized in any other form or if we wish to be able to display, and perhaps play, a specific version of the computer game.

The same analysis can be applied to digital photographs. With traditional photographs, one would say that altering the image that

had been captured on film was contrary to preserving it. But there are several types of digital photographs where the possibilities of displaying different images of the same picture are valuable. For example, a traditional chest X-ray produced three pieces of film, and, therefore, three fixed images. But a computerized axial tomography (CAT) scan of the chest can produce scores of different images, making it a more flexible and incisive tool for diagnosis. How should CAT scans be preserved? It depends on our conception or model of what a CAT scan is. If we wanted to preserve the richest source of data about the state of a particular person's body at a given time, we would have to preserve the CAT scan as an instance of a specific type of technology. But if we needed to preserve a record of the specific image that was the basis for a diagnosis or treatment decision, we would have to preserve it as a specific image whose visual appearance remains invariant over time. If the first case, we must preserve CAT scanning technology, or at least that portion of it necessary to produce different images from the stored bit file. It is at least worth considering, in the latter case, that the best preservation method, taking feasibility and sustainability into account, would be to output the image on archival quality photographic film.

Here, in the practical context of selecting preservation methods, we see the operational importance of the principle articulated in discussing the authenticity of preserved objects: we can determine what is needed for preservation only on the basis of a specific concept or definition of the essential characteristics of the object to be preserved. The intended use of the preserved objects is enabled by the articulation of the essential characteristics of those objects, and that articulation enables us not only to evaluate the appropriateness of specific preservation methods but also to determine how they should be applied in any case. Applying the criterion of appropriateness, we can align various preservation methods across the spectrum of "preserve technology"—"preserve objects."

More than a Spectrum: A Two-Way Grid

For any institution that intends or needs to preserve digital information objects, selection of preservation methods involves another dimension: the range of applicability of the methods with respect to the quantity and variety of objects to be preserved. Preservation methods vary greatly in terms of their applicability. Some methods apply only to specific hardware or software platforms, others only to individual data types. Still others are very general, applicable to an open-ended variety and quantity of digital objects. The range of applicability is another basis for evaluating preservation methods. Organizations that need to preserve only a limited variety of objects can select methods that are optimal for those objects. In contrast, organizations responsible for preserving a wide variety must select methods with broad applicability. Combining the two discriminants of appropriateness for preservation objectives and range of applicabili-

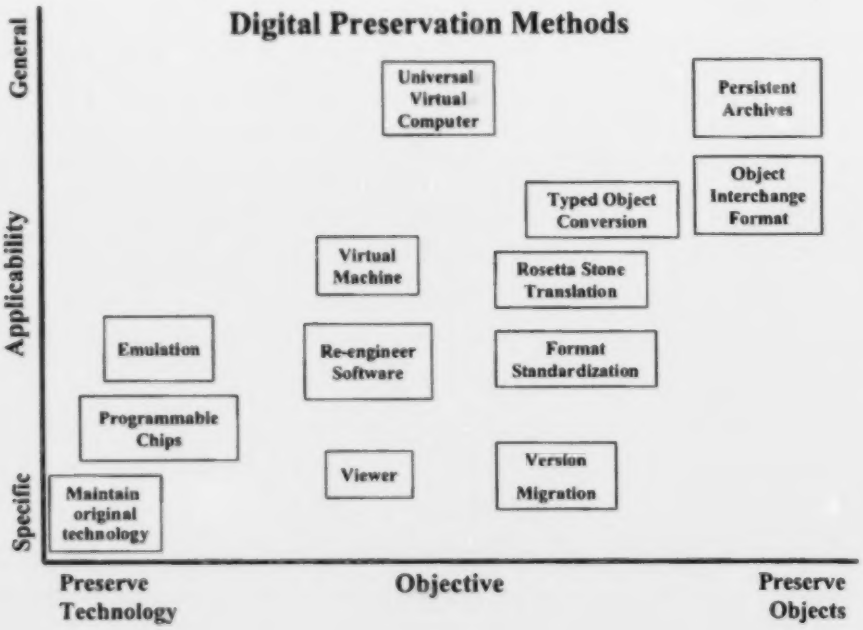
ty defines a two-dimensional grid in which we can place different preservation methods and enrich our ability to evaluate them.

Figure 4 shows this grid, with a number of different methods positioned in it. Two general remarks about the methods displayed in this grid are in order. On the one hand, the methods included in it do not include all those that have been proposed or tried for digital preservation. In particular, methods that focus on metadata are not included. Rather, the emphasis is on showing a variety of ways of overcoming technological obsolescence. Even here, the cases included are not exhaustive; they are only illustrative of the range of possibilities. On the other hand, some methods are included that have not been explicitly or prominently mentioned as preservation methods. There is a triple purpose for this. The first purpose is to show the robustness of the grid as a framework for characterizing and evaluating preservation methods. The second is to emphasize that those of us who are concerned with digital preservation need to be open to the possibilities that IT is constantly creating. The third purpose is to reflect the fact that, in the digital environment, preservation is not limited to transmitting digital information over time. The same factors are in play in transmitting digital information across boundaries in space, technology, and institutions. Therefore, methods developed to enable reliable and authentic transmission across one of these types of boundaries can be applicable across others (Thibodeau 1997).

Sorting IT Out

Discussions of digital preservation over the last several years have focused on two techniques: emulation and migration. Emulation strives to maintain the ability to execute the software needed to pro-

Fig. 4. Digital Preservation Methods



cess data stored in its "original" encodings, whereas migration changes the encodings over time so that we can access the preserved objects using state-of-the-art software in the future. Taking a broader perspective, IT and computer science are offering an increasing variety of methods that might be useful for long-term preservation. These possibilities do not fit nicely into the simple bifurcation of emulation versus migration. We can position candidate methods across the preservation spectrum according to the following principles:

- On the "preserve technology" end of the spectrum, methods that attempt to keep data in specific logical or physical formats and to use technology originally associated with those formats to access the data and reproduce the objects.
- In the middle of the spectrum, methods that migrate data formats as technology changes, enabling use of state-of-the-art technology for discovery, access, and reproduction.
- On the "preserve objects" end of the spectrum, methods that focus on preserving essential characteristics of objects that are defined explicitly and independently of specific hardware or software.

There are various ways one can go about all these options. For example, if we focus on the "preserve technology" end, we start with maintaining original technology, an approach that will work for some *limited* time. Even for preservation purposes, it can be argued that this approach is often the only one that can be used.

Preserving Technology: The Numbers Add Up, and Then Some

The starting point for all digital preservation is the technology and data formats used to create and store the objects. Digital information objects can be preserved using this "original" technology for 5 to 10 years, but eventually the hardware, software, and formats become obsolete. Trying to preserve specific hardware and software becomes increasingly difficult and expensive over time, with both factors compounded by the variety of artifacts that need to be preserved. Over the long term, keeping original technology is not practicable and may not be feasible.

Enter the Emulator

Various approaches can be used to simplify the problem while still keeping data in their original encodings. The best-known approach is emulation. Emulation uses a special type of software, called an emulator, to translate instructions from original software to execute on new platforms. The old software is said to run "in emulation" on newer platforms. This method attempts to simplify digital preservation by eliminating the need to keep old hardware working. Emulators could work at different levels. They could be designed to translate application software to run on new operating systems, or they could translate old operating system commands to run on new operating systems. The latter approach is simpler in that the former

would require a different emulator for every application, and potentially for every version of an application, while the latter should enable all applications that run on a given version of an operating system to execute using the same emulator.

While proponents of emulation argue that it is better than migration because at every data migration there is a risk of change, emulation entails a form of migration. Emulators themselves become obsolete; therefore, it becomes necessary either to replace the old emulator with a new one or to create a new emulator that allows the old emulator to work on new platforms. In fact, if you get into an emulation strategy, you have bought into a migration strategy. Either strategy adds complexity over time.

Emulation is founded on the principle that all computers are Turing machines and that any command that can run on one Turing machine can run on any other Turing machine. There is, however, evidence that this principle breaks down at an empirical level. For example, basic differences such as different numbers of registers or different interrupt schemes make emulation unreliable, if not impossible (IEEE 2001).

Reincarnation for Old Machines

Another technique that keeps old software running takes the opposite approach from emulation: it relies on a special type of hardware, rather than software emulators. It does this by re-creating an old computer on a configurable chip. An entire computer system could be reincarnated by being programmed on a new, configurable chip. The configurable chip constitutes a single point of failure, but that can readily be offset. If the chip begins to fail or becomes obsolete, the old system could simply be programmed on a newer chip. Intuitively, configurable chips seem like a simpler approach than emulation.

Compound Disinterest

While emulation and configurable chips take opposite directions, they present some common problems. First, current technology is not perfect. There are anomalies and bugs. Any preservation strategy that relies on specific software is carrying all the problems associated with those products into the future. Not all these problems get fixed. For example, it is not always possible to figure out what causes a problem such as a general protection fault, because there are too many variables involved. Furthermore, fixes can increase the complexity of preservation strategies that rely on keeping old software running, because they increase the number of versions of software that are released. Logically, if the authenticity of digital information depends on preserving original data formats and using them with the original software, each format should be processed with the version of the software used to produce it.

Software defects aside, the combinatorics entailed by strategies that involve preserving ever-increasing varieties of data formats, application software, and operating systems are frightening. With new versions being released every 18–24 months, over 25-years or longer,

one would need to support thousands of combinations of applications, utilities, operating systems, and formats.

The viability of these strategies gets much more complex when the focus shifts from a single system to combinations of systems, which is the norm today. Emulation and programmable chips might be viable strategies if all we had to cope with were the products of desktop PCs, but not in today's world, where the objects to be preserved often involve a diverse palette of technologies, such as various client-server applications where the servers use different operating systems, distributed applications running on heterogeneous platforms, and virtual machines such as Java. Providing technical support for operations of such a daunting variety of makes, models, and versions may be neither feasible nor affordable, because you would have to get all these applications running in emulation at the same time.

Complexity also increases in the case of collections of documents accumulated over time. Most government records, for example, are accumulated over many years, often many decades. Following the most fundamental principles of archival science—respect for provenance and for original order—we cannot segregate records according to their digital formats. We must preserve and provide access to aggregates of records established by their creators. Under a strategy of preserving technology, doing research in such series would entail using all the different software products used to produce the records.

Even if it were technically and financially possible to keep the technologies operative, staffing a help desk to support end users is inconceivable, especially since most users in the future will never have encountered—not to mention learned how to use—most of the products they will need to access the preserved information objects. Even if it were possible to provide adequate support to a user perusing, for example, a single case file accumulated over 20 years, it is not obvious that this would be deemed an acceptable level of support, because it would cut users off from the possibility of using more advanced technologies for discovery, delivery, and analysis.

Scenarios pegged on preserving specific technology, maintaining the links between specific software and specific data formats, run counter to the major direction of information technology. E-commerce and e-government require that the information objects created and used in these activities be transportable among the parties involved, independent of the hardware and the software each party uses at any time. Neither e-commerce nor e-government would be possible if the necessary information had to be accessed in original formats using obsolete technologies. Preserve technology strategies will depend on niche technologies and cannot expect widespread support in the IT market.

In this approach, one also encounters some interesting issues of intellectual property rights—not only the usual issues of copyright but also the ownership that the software companies assert over their formats even when they do not own the content.

A View Toward Further Simplification

Various software-engineering methods provide simpler ways of keeping obsolete formats accessible by concentrating on specific requirements.

One such method focuses on documents, a class of objects in which the functionality that has to be preserved is simply the ability to present them visually on a screen or printed page. For such objects, the only specific software needed for preservation is software that reliably renders the content with its original look and feel. This approach is being used in the Victorian Electronic Records System (VERS) developed for the Public Record Office of the State of Victoria, Australia. The system stipulates converting documents created in desktop computing environments to Adobe's PDF format. Instead of attempting to run versions of Acrobat reader for PDF indefinitely in the future, the VERS project conducted an experiment to demonstrate that it is possible to construct a viewer from the published specifications of the PDF format. The VERS approach embodies a combination of format migration, in that the various formats in which records are originally created must be translated to PDF with software reengineering. Similar approaches could be applied to other data types whose essential functionality is presentation in page image.

Finding Virtue in Abstraction

Another application of software engineering involves developing virtual machines that can execute essential functions on a variety of platforms. The Java language is an example of a virtual machine, although it was not developed for purposes of preservation. The virtual machine approach avoids the need for emulator software by providing required functionality in a virtual machine that, in principle, can be implemented on a great variety of computing platforms indefinitely into the future. Raymond Lorie of the IBM-Almaden Research Center has launched an effort to develop a Universal Virtual Computer (UVC) that would provide essential functionality for an unlimited variety of data types. Following this strategy, objects would be preserved in their original formats, along with the rules for encoding and decoding data in those formats. The rules are written in a machine language that is completely and unambiguously specified. The language is so simple that it can be interpreted to run on any computer in the future. When the UVC program executes, the preserved data are interpreted according to a logical schema for the appropriate data type and output, and each data element bears a semantic tag defined in the logical schema. This approach avoids much of the complexity of emulation and configurable chips, but there are some trade-offs. The UVC only provides a limited set of basic functions. It also sacrifices performance: software that can run on any platform is not optimized for any one of them (Lorie 2000).

Accepting Change: Migration Strategies

In the middle of the spectrum fall data migration approaches that abandon the effort to keep old technology working or to create substitutes that emulate or imitate it. Instead, these approaches rely on changing the digital encoding of the objects to be preserved to make it possible to access those objects using state-of-the-art technology after the original hardware and software become obsolete. There are a variety of migration strategies.

Simple Version Migration

The most direct path for format migration, and one used very commonly, is simple version migration within the same family of products or data types. Successive versions of given formats, such as Corel WordPerfect's WPD or Microsoft Excel's XLS, define linear migration paths for files stored in those formats. Software vendors usually supply conversion routines that enable newer versions of their product to read older versions of the data format and save them in the current version.

Version migration sets up a chain that must be extended over time, because every format will eventually become obsolete. One problem with this approach is that using more recent versions of software, even with the original formats, may present the preserved documents with characteristics they did not, and perhaps could not, have had. For example, any document created with a word processor in the early 1990s, before "WYSIWYG" display was available, would have appeared on screen with a black background and green letters. If one were to open such a document with a word processor today, it would look much like this document.

Software vendors control this process of version migration. Their conversion utilities are designed to migrate data types and do not provide for explicit or specific control according to attributes defined at the conceptual level. Each successive migration will accumulate any alterations introduced previously. Another potential problem is that over time, product lines, and the migration path, may be terminated.

Format Standardization

An alternative to the uncertainties of version migration is format standardization, whereby a variety of data types are transformed to a single, standard type. For example, a textual document, such as a WordPerfect document, could be reduced to plain ASCII. Obviously, there would be some loss if font, type size, and formatting were significant. But this conversion is eminently practicable, and it would be appropriate in cases where the essential characteristics to be preserved are the textual content and the grammatical structure. Where typeface and font attribute are important, richer formats, such as PDF or RTF, could be adopted as standards. The low common denominator provides a high guarantee that the format will be successful, at least for preserving appearance. For types of objects where visual presentation is essential, bit-mapped page images and hard

copy might be acceptable: 100 years from now, IT systems will be able to read microfilm. In fact, according to companies such as Kodak and Fuji, it can be done today.

For socioeconomic and other data sets created to enable a variety of analyses, the data structure can often be preserved in a canonical form, such as arrays or relational tables, independently of specific software. Such formats are either simple enough or so unambiguously defined that it is reasonable to assume that information systems in the future will be able to implement the structures and process the data appropriately.

In principle, the standard format should be a superclass of the original data types—one that embodies all essential attributes and methods of the original formats. This is not necessarily the case, so there may be significant changes in standardization, just as with version migration. Moreover, standards themselves evolve and become obsolete. So, except for the simplest formats, there is a likely need for repeated migrations from one standard format to another, with consequent accumulation of changes.

Typed Object Model Conversion

Another approach to migrating data formats into the future is Typed Object Model (TOM) Conversion. The TOM approach starts out with the recognition that all digital data things are objects, that is, they have specified attributes, specified methods or operations, and specific semantics. All digital objects belong to one or another type of digital object, where "type" is defined by given values of attributes, methods, or semantics for that class of objects. A Microsoft Word 6 document, for example, is a type of digital object defined by its logical encoding. An e-mail is a type of digital object defined, at the conceptual and logical levels, by essential data elements, e.g., "To," "From," "Subject," or "Date."

Any digital object is a byte sequence and has a format, i.e., a specified encoding of that object for its type. Byte sequences can be converted from one format to another, as shown in the earlier example of this document encoded in Microsoft Word and PDF formats. But within that range of possible conversion, the essential properties of a type or class of objects define "respectful conversions," that is, conversions whose result cannot be distinguished when viewed for an interface of that type. The content and appearance of the document in this example remains identical whether it is stored as a Word or PDF file; therefore, conversion between those two formats is respectful for classes of objects whose essential properties are content and appearance (Wing and Ockerbloom 2000). There is a TOM conversion available online that is capable of doing respectful conversions of user submitted files in some 200 formats.

Rosetta Stones Translation

Another migration approach under development is called Rosetta Stones. Arcot Rajasekar of the San Diego Supercomputer Center is developing this approach. Like TOM, this approach starts with data

type s, but rather than articulating the essential properties of each type, it constructs a representative sample of objects of that type. It adds a parallel sample of the same objects in another, fully specified type, and retains both. For example, if one wanted to preserve textual documents that had been created in WordPerfect 6, one would create a sample of files in version 6 of the WPD format that embodies all the significant features of this format. Then one would duplicate each of the documents in this sample in another format that might be human-readable computer output microfilm (COM) or paper, because we know that we will always be able to read in those human-readable versions. This second sample constitutes a reference set, like the Greek in the original Rosetta Stone. The triad of samples in the original data type, the reference set, and the target type constitutes a digital Rosetta Stone from which rules for translating from the original to the target encoding can be derived.

Given the reference sample—e.g., the printed version of documents—and the rules for encoding in a target format that is current at any time in the future, we can create a third version of the sample in the target format. By comparing the target sample with the original sample, we can deduce the rules for translating from the original to the target format and apply these rules to convert preserved documents from the original to the target format. This approach avoids the need for repeated migrations over time. Even though the target formats can be expected to become obsolete, migration to subsequent formats will be from the original format, not from the earlier migration. Important to the success of this approach is the ability to construct a parallel sample in a well-characterized and highly durable type. It is not evident that it will be possible to do this for all data types, especially more complex types that do not have analog equivalents, but research on this approach is relatively recent.

Object Interchange Format

Another approach enables migration through an object interchange format defined at the conceptual level. This type of approach is being widely adopted for e-commerce and e-government where participants in a process or activity have their own internal systems, which cannot readily interact with systems in other organizations. Rather than trying to make the systems directly interoperable, developers are focusing on the information objects that need to be exchanged to do business or otherwise interact. These objects are formally specified according to essential characteristics at the conceptual level, and those specifications are articulated in logical models. The logical models or schema define interchange formats. To collaborate or interact, the systems on each side of a transaction need to be able to export information in the interchange format and to import objects in this format from other systems. While it was designed for internal markup of documents, the XML family of standards has emerged as a major vehicle for exchange of digital information between and among different platforms.

A significant example of this approach concerns financial reports. There are several types of financial reports that essentially all corporations produce and share with their business partners and with government agencies around the world. The extensible business reporting language (XBRL) is an initiative to enable exchange of these reports, regardless of the characteristics of systems used either to produce or receive the reports. The initiative comprises major professional organizations of accountants from the United States, Canada, the United Kingdom, Australia, and several non-English-speaking countries, major accounting firms, major corporations, IT companies, and government agencies. XBRL defines a single XML schema that covers all standard financial reports. The schema defines an interchange format. Any system that can export and import data in that format can exchange financial reports and data with any other system with XBRL I/O capability, regardless of the hardware or software used in either case. At the logical level, the XBRL schema is impressively simple. That simplicity is enabled by an extensive ontology of accounting terms at the conceptual level. This approach is obviously driven by short-term business needs, but a method that allows reliable exchange of important financial data across heterogeneous computing platforms around the world can probably facilitate transmission of information over generations of technology. Given that XML schemas and tags are constructed using plain ASCII and can be interpreted by humans, it is likely that future computer systems will be able process them correctly. Thus, the object interchange method can become a preservation method simply by retaining the objects in the interchange format and, on an as-needed basis, building interpreters to enable target systems in the future to import objects in such formats.

To some extent, object interchange formats have the same purpose as do samples in well-known data types in the Rosetta Stones method: they serve as a bridge between heterogeneous systems and data types. While the Rosetta Stones method is more generic, object interchange specifications have a significant advantage in that the essential properties of the objects are defined by experts who have substantial knowledge of their creation and use. Thus, unlike all the other approaches considered so far, object interchange formats embed domain knowledge in the transmission of information objects across space, time, and technologies. The object interchange model lies close to the "preserve objects" end of the preservation spectrum. It could be said to lie midway between specific and general in its applicability because it provides a single method that potentially could be applied to a great variety of objects and data types, but addresses only the persistence of content and form across technological boundaries.

Preserving Objects: Persistent Archives

A promising approach, persistent archives, has been articulated over the last four years, primarily at the San Diego Supercomputer Center in research sponsored by the Defense Advanced Research Projects

Agency, the National Science Foundation, and the National Archives and Records Administration. It has many elements in common with other approaches described in this paper, but it is also markedly different than these other strategies. Like the UVC, it relies on a high level of abstraction to achieve very broad applicability. Like TOM and Rosetta Stones, it addresses the specific characteristics of logical data types. Like object interchange formats and the UVC, it tags objects to ensure the persistence of syntactic, semantic, and presentation elements. Like migration, it transforms the logical encoding of objects, but unlike migration, the transformations are controlled not by target encodings into which objects will be transformed but by the explicitly defined characteristics of the objects themselves. It implements a highly standardized approach, but unlike migration to standard format, it does not standardize on logical data types, but at a higher level of abstraction: on the method used to express important properties, such as context, structure, semantics, and presentation.

The most important difference between persistent archives and the other approaches described is that the former strategy is comprehensive. It is based on an information management architecture that not only addresses the problem of obsolescence but also provides the functionality required for long-term preservation, as stipulated in the OAIS standard. Furthermore, it provides a coherent means of addressing the physical, logical, and conceptual properties of the objects being preserved through the data, information, and knowledge levels of the architecture. Persistence is achieved through two basic routes: one involving the objects themselves, the other the architecture. Objects are preserved in persistent object format, which is relatively immune to the continuing evolution of IT. The architecture enables any component of hardware or software to be replaced with minimum impact on the archival system as a whole. The architecture is notional. It does not prescribe a specific implementation.

The cornerstone of the persistent archives approach is the articulation of the essential characteristics of the objects to be preserved—collections as well as individual objects—in a manner that is independent of any specific hardware or software. This articulation is expressed at the data level by tags that identify every byte sequence that must be controlled to ensure preservation. In effect, tags delimit atomic preservation units in physical storage. The granularity of these data units can vary greatly, depending on requirements articulated at the information and knowledge levels. Every tag is linked to one or more higher-level constructs, such as data models, data element definitions, document type definitions, and style sheets defined at the information level, and ontologies, taxonomies, thesauri, topic maps, rules, and textbooks at the knowledge level. In research tests on a wide variety of data types, conceptual objects, and collections, it has been shown that simple, persistent ASCII tags can be defined to identify, characterize, and control all data units. The research has shown that XML is currently the best method for tagging and articulating requirements at the information level and, to some extent, at the knowledge level; however, it would be wrong to conclude that

persistent archives are based or dependent on XML. Rather, persistent archives currently use XML, but there is nothing in the architecture that would preclude using other implementation methods should they prove superior.

The architecture is structured to execute the three basic processes required in the Open Archival Information System (OAIS) standard: *ingest*, for bringing objects into the system; *management*, for retaining them over time; and *access*, for disseminating them to consumers. In ingest, objects in obsolescent formats are transformed into persistent format, through parsing and tagging of data units as described earlier, or, if they are already in persistent format, by verifying that fact at the data, information, and knowledge levels. Over time, data units are maintained in storage, and the metadata and domain knowledge that are necessary to retrieve, use, and understand the data are maintained in models, dictionaries, and knowledge bases. When access to a preserved object is desired, the data are retrieved from storage and the object is materialized in a target technology current at the time. This materialization requires translating from the persistent form to the native form of the target technology. If the three basic processes are conceived as columns and the three levels (data, information, knowledge) as rows, the persistent archives architecture can be depicted in a 3-by-3 grid (Moore et al. 2000).

The persistent archives architecture is independent of the technology infrastructure in which it is implemented at any time. It achieves this independence through loose coupling of its basic building blocks, using software mediators to link each pair of adjacent blocks. Interactions are between adjacent blocks vertically and horizontally, but not diagonally. Over time, as the components used to implement any block are updated, there is no need to change any of the other blocks, only the mediators.

Conclusion: The Open End

There is an inherent paradox in digital preservation. On the one hand, it aims to deliver the past to the future in an unaltered, authentic state. On the other hand, doing so inevitably requires some alteration. All the methods described in this paper entail altering or replacing hardware, software, or data, and sometimes more than one of these. This paradox is compounded by the fact that in the future, as today, people will want to use the best available technology—or at least technologies they know how to use—for discovery, retrieval, processing, and delivery of preserved information. There is a danger that to the degree that preservation solutions keep things unaltered they will create barriers to satisfying this basic user requirement. Adding to this the recognition that the problem of digital preservation is not static, that it will continue to evolve as information technology and its application in the production of valuable information change, reinforces the paradox to the point that any solution to the challenge of digital preservation must be inherently evolutionary. If the preservation solution cannot grow

and adapt to continuing changes in the nature of the problem and continuing escalation of user demands, the "solution" itself will in short order become part of the problem; that is, it will itself become obsolete.

This paradox can be resolved only through the elaboration of a basic conceptual framework for digital preservation—a framework that allows us to identify and analyze all that is involved in the process of digital preservation and to understand how different facets of that process affect each other. Fortunately, such a framework has been articulated over the last few years and has become an international standard. It is the OAIS reference model. While the OAIS model was developed for the space science community, its articulation was, from the beginning, both international and multidisciplinary. As a result, the model has broad applicability. The OAIS model provides a frame of reference in which we can balance the need for preserving digital objects unaltered and the need to keep pace with changing IT, both to encompass new classes of digital objects and to capitalize on technological advances to improve preservation services (ISO 2002).

However, the OAIS model is too generalized to suffice for implementation. It needs to be refined and extended to be useful in specific domains. One example of such refinement has been articulated for the domain of records. The International research on Permanent Authentic Records in Electronic Records (InterPARES) project is a multinational, multidiscipline research collaboration whose name reflects its basic objective. To fine-tune the OAIS framework for the specific goal of preserving authentic records, the InterPARES Project developed a formal Integrated DEFinition (IDEF) process model for what is required to preserve authentic digital records. This "Preserve Electronic Records" model retains the functions of an OAIS but adds specific archival requirements and archival knowledge. Archival requirements act as specific controls on the preservation process, and archival knowledge was the basis for further refinement of the preservation process. In turn, the process of developing the archival model led to advances in archival knowledge; specifically, to clarification of the characteristics of electronic records at the physical, logical, and conceptual levels, and to improvements in our understanding of what it means to preserve electronic records. The InterPARES Preserve Electronic Records model includes specific paths for accommodating new classes of electronic records over time and for taking advantage of improvements in IT (Preservation Task Force in press).

The InterPARES model illustrates how, starting from the OAIS reference model, one can construct an open-ended approach to digital preservation and effectively address the paradoxical challenge of digital preservation. This case can serve as an example for other domains. There is undeniably a pressing need for technological methods to address the challenge of digital preservation. There is a more basic need for an appropriate method of evaluating alternative methods, such as the two-way grid described in this paper. Fi-

nally, there is an overriding need to select and implement preservation methods in an open-ended system capable of evolving in response to changing needs and demands.

References

All URLs were valid as of July 10, 2002.

Institute of Electrical and Electronics Engineers, Inc. (IEEE). June 2001. Transactions on Computers—Special Issue on Dynamic Optimization.

International Standards Organization (ISO). 2002. Open Archival Information System—Reference Model. The draft international standard is available at http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html.

InterPARES Project Preservation Task Force. 2001. How to Preserve Electronic Records. Available at www.interpares.org.

InterPARES Project Preservation Task Force. In press. Report of the Preservation Task Force. A preliminary version of the report is available at www.interpares.org.

Jefferson, Thomas. 1801. Note from Thomas Jefferson regarding the disposition of his Presidential papers, December 29, 1801. Washington, D.C.: National Archives, General Records of the Department of State. RG 59. Available at http://nara.gov/education/teaching/archives/tjletter.html.

Lorie, Raymond A. 2000. The Long-Term Preservation of Digital Information. Available at http://www.si.umich.edu/CAMILEON/Emulation%20papers%20and%20publications/Lorie.pdf.

Lynch, Clifford. 2000. Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust. In *Authenticity in a Digital Environment*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/pubs/abstract/pub92abst.html.

Moore, Reagan et al. 2000. Collection-Based Persistent Digital Archives—Part 1. *D-Lib Magazine* 6(3). Available at http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.

Rajasekar, Arcot, Reagan Moore, and Michael Wan. Syntactic and Semantic Replicas: Rosetta Stones for Long-Term Digital Preservation. Unpublished manuscript.

Rothenberg, Jeff. 2000. Preserving Authentic Digital Information. In *Authenticity in a Digital Environment*. Washington, D.C.: Council on Library and Information Resources. Available at http://www.clir.org/pubs/abstract/pub92abst.html.

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* Washington, D.C.: Commission on Preservation and Access, and Mountain View, Calif.: Research Libraries Group. Available at http://www.rlg.org/ArchTF/.

Thibodeau. Kenneth. 1997. Boundaries and Transformations: An Object-Oriented Strategy for Preservation of Electronic Records. European Commission. In *INSAR Supplement II: the Proceedings of the DLM-Forum on Electronic Records*. Luxembourg: Office for Official Publications of the European Communities.

Wing, Jeannette M., and John Ockerbloom. 2000. Respectful Type Converters. *IEEE Transactions on Software Engineering* 26(7): 579-93.

Web sites noted in this paper:

Extensible Business Reporting Language (XBRL). Available at http://www.xbrl.org/.

Typed Object Model (TOM). Available at http://tom.cs.cmu.edu/intro.html.

Victorian Electronic Records Strategy (VERS). Available at http://www.prov.vic.gov.au/vers/.

The Digital Preservation Research Agenda

Margaret Hedstrom

Research in digital archiving and long-term preservation is an increasingly popular topic for discussion. Although there have been many calls for research on long-term preservation of digital objects over the past decade, the present environment is especially conducive to defining a research agenda and developing effective research programs.

This paper focuses on three aspects of digital preservation research. It begins with a discussion of needs and opportunities that distinguish current efforts from previous attempts to organize research programs on digital preservation. It then describes some potential frameworks for research. The paper concludes with some recommendations for research programs that are methodologically and conceptually sound as well as useful to a broad community.

Current Needs and Opportunities

Two aspects of the current environment—need and opportunity—provide reason for optimism about the prospects for digital preservation research. Whereas those engaged in previous attempts to articulate digital preservation research issues were forced to spend a great deal of effort simply defining the problem, we now have a firm foundation on which to build research programs. During the past decade or so, libraries, archives, scientific data centers, government agencies, corporations, and private individuals have built significant collections of digital content. Many wonderful collections have been built through retrospective conversion, and a growing amount of born-digital content has been captured, in some way, by widely dispersed organizations. The need for preservation research is no longer a hypothetical question based on the premise that if we create valuable digital content, then someone will have to be concerned with its lon-

gevity. The need for preservation is real, and the absence of a preservation strategy is increasingly acknowledged as an obstacle to the full realization of a digital future. Numerous research projects, some of which are still under way, provide a foundation for identifying which approaches seem promising, for honing research methodologies, and for demonstrating the benefits of sound research. The collaboration between the National Archives and Records Administration (NARA) and the San Diego Supercomputer Center, and the Mellon Foundation-funded research projects on e-journal archives are two examples of at least a dozen such projects.

Significant opportunities exist for designing research programs that will advance basic knowledge and will also provide practical tools and solutions for libraries and archives, organizations with significant digital assets, and even for private citizens who are concerned about their own personal digital archives. Three such opportunities deserve mention. First, as part of the Library of Congress (LC)-led initiative to develop a National Digital Information Infrastructure and Preservation Program, the National Science Foundation (NSF) and LC have been working together to develop a research agenda for long-term digital archiving and preservation. A workshop held on April 12 and 13, 2002, focused on the research challenges in digital archiving and on building a national infrastructure for long-term preservation of digital information. Attending this workshop were 50 participants from industry, academia, and the government. The session engendered a sophisticated cross-fertilization of ideas among researchers in archives, information science, digital libraries, and computer science. The report from the workshop, which will be published this summer, will present a research agenda that funding agencies will use to mobilize resources for sponsored research universities.[1]

A related, and potentially more significant, development is the recent release of the draft report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure (NSF 2002). This panel, chaired by Daniel E. Atkins, investigated the types of investments that NSF and other organizations need to make to create an infrastructure for advanced research in science and engineering. A cornerstone of the panel's vision for a cyberinfrastructure is a network of knowledge-management institutions for collection building and curation of data, information, literature, and digital objects. The draft report recommends support for 50 to 100 data repositories that are grounded in the domain sciences where NSF funds research. The annual cost of this effort is estimated at $140 million.

A third area of opportunity is international collaboration. The presence of several participants from overseas at this symposium and the presentations by Titia van der Werf and Colin Webb indicate

[1] Additional information about this workshop is available at: www.si.umich.edu/digarch/. Support for this workshop was provided by National Science Foundation award #021469. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

the global aspects of this issue. Researchers and institutions in the United States that are building digital repositories benefit tremendously from work under way in Australia, Germany, the Netherlands, the Nordic countries, the United Kingdom, and elsewhere. Several participants in this symposium are members of the joint Working Group on Digital Preservation Research, funded by the NSF and DELOS, a European Union (EU) initiative to promote excellence in digital libraries in the EU. This is one of several opportunities to coordinate research internationally.

Frameworks for Research and Practical Applications

Despite these opportunities, a tension remains between immediate needs for solutions and the potential lag in transferring research results into practical applications. The initial conclusions of the NSF workshop indicate a consensus on some conceptual models for digital archiving. This consensus underlies the models presented by Kenneth Thibodeau of NARA, which separate physical storage from logical interpretation and distinguish data management from knowledge management. The concept of organizing the digital preservation challenge into a series of components, or layers, in a model architecture provides a basis for distributing responsibility among various types of institutions. Moreover, elements of the basic architecture can change over time without requiring that an archival system be redesigned. There is a strong consensus that this framework is an important step forward, even though there is also agreement that there is no single answer to all digital preservation problems. We need a spectrum of solutions in terms of scale, format types, and institutional responsibilities.

There is a clear sense that effective practices exist for certain types of static digital objects. In those cases where there is a strong basis of knowledge that organizations can use to move forward with implementation, we need to teach practitioners about these methods and practices. As a community, we may need to agree that while these practices are not perfect, they are effective enough to serve as a basis for moving forward with implementation. Such implementations could readily occur in the area of reformatting and conversion of traditional materials for certain formats. We have excellent guidelines and best practices for print documents and images that are oriented to building collections with qualities that make them more easily preserved over the long run. Anne Kenney and Oya Reiger address these practices in *Moving Theory Into Practice* (2000). The Library of Congress provides sound guidance through the technical requirements developed by its American Memory Program.[2] Standards and guidelines endorsed by the Digital Library Federation, such as the Framework for Building Good Digital Collections developed by the Institute of Museums and Library Services, also repre-

[2] Technical guidance is available at: http://memory.loc.gov/ammem/ftpfiles.html.

sent community-based best practices.[3] Any organization that is building digital collections should follow these guidelines, and funding agencies should require that all sponsored projects conform to them. In the area of static born-digital documents, some models are emerging for converting materials from proprietary formats into extensible markup language (XML). On the other hand, complex and dynamic objects present a significant challenge that requires considerable research. We also have much to do before reaching a consensus on best practices for video, film, recorded sound, and multimedia.

Preservation research can draw on related research in computer science and information science. Digital preservation shares many requirements with well-designed information systems, such as security, authentication, robust models for representation, and sophisticated information retrieval mechanisms. By adapting related research to meet some digital preservation challenges, we will be able to focus on the unique problems of long-term preservation. Participants in the April workshop on digital preservation challenges discussed some of the problems that are unique to long-term preservation.

One unique aspect of preservation is its concern with the long term, where "long term" does not necessarily mean generations or centuries. It may simply mean long enough to be concerned about the obsolescence of technology. In this area, preservation requirements may exceed what information technology vendors typically provide. When long-term preservation spans several decades, generations, or centuries, the threat of interrupted management of digital objects becomes critical. Digital objects cannot be left in an obsolete format and then turned over to a repository after a long period of neglect. This challenge is as much a social and institutional problem as it is a technical one, because for long-term preservation, we rely on institutions that go through changes in direction, purpose, management, and funding.

Considerable research is needed to develop funding and business models for repositories that assume preservation responsibilities. Repositories may be expected to preserve digital resources even though their utility may not become apparent until well into the future and even though the future users are not yet born. Over the long term, new communities of users will emerge with needs and expectations that differ from those of the communities that created the digital content. The challenge of developing economic models for the value and costs of archiving over the long term deserves an entire meeting or conference.

Another factor that distinguishes digital preservation research from many other types of research is the difficulty of knowing whether or not we have solved the problems. We may know when we have failed, but we may not be alive to know whether we have succeeded. This problem requires some challenging thinking about success measures and evaluation criteria.

[3] These guidelines are available at http://www.diglib.org/standards.htm.

Methodologies for Research and Knowledge Transfer

How we carry out research may be as important as the topics we choose to investigate. There are some frameworks that can move the field of digital preservation research forward. One recommendation is to disaggregate digital preservation research issues into manageable problems. The principles for this disaggregation are not yet established, but one place to start is by distinguishing between preservation of converted materials and born-digital content, between static and dynamic objects, among different formats, and between different producer and user communities. So far, very small investments have been made in research on a very large problem. If we can develop frameworks that allow people to apply their specialized knowledge and skills to specific problems, we can move forward.

One informative concept comes from *Pasteur's Quadrant* by Donald Stokes (1997). *Pasteur's Quadrant* breaks down the tired dichotomy between basic and applied research. Using a four-cell matrix, Stokes provides a dynamic model that allows for considerations of use to inform a basic quest for understanding. He uses the example of Niels Bohr, who was seeking basic understanding; Thomas Edison, who was trying to build something useful; and Louis Pasteur, who is in the quadrant where use, demand, and interest intersect with a quest for finding basic answers.

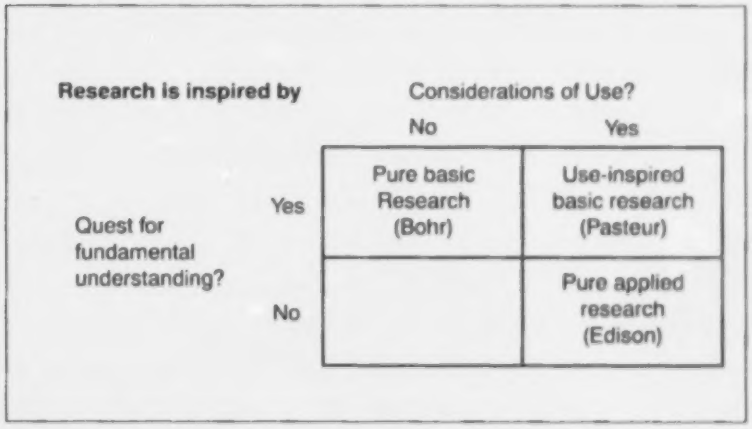


Fig. 1. Quadrant Model of Scientific Research

Source: Stokes 1997, 74. Reprinted with permission by the Brookings Institution Press.

To the extent that we can design research projects that fit into that quadrant of "use-inspired basic research," we will benefit both from what the academic research community has to offer and from the interesting questions that practitioners present.

Potential research methodologies cover a spectrum—from theory building to exploratory research, simulations, and experiments. One difference between digital preservation research and research on preserving physical objects is that we can make copies of bits or objects and experiment with them. We can run digital objects through a number of processes and get observable and measurable results. Such experiments would allow researchers to compare the results of different preservation strategies in terms of effectiveness, cost, and user acceptance. For example, a series of experiments comparing em-

ulation and migration would allow researchers to conclude that for a particular type of digital object, an emulation approach preserves *these* specific properties, has *these* complications, and would cost *this amount* of money, whereas a migration approach to the same material over three format conversions has *these* specific consequences and costs *this* much. We need more concrete evidence and an empirical basis for evaluating different preservation strategies and for deciding which strategy is most appropriate for particular types of resources.

A considerable amount of enthusiasm is building around the idea of creating test beds where a designer or researcher—or more likely a large team of researchers—creates a prototype environment that has metrics that will make it possible to measure the effectiveness of different strategies. The work of the San Diego Supercomputer Center falls under the definition of a test bed, where libraries, archives, and organizations can bring real collections and problems as experimental data sources. A test bed also involves a feedback loop among the people with collections to manage and the people designing and running test beds. Knowledge transfer and technology transfer remain significant challenges. Researchers can do wonderful things in the lab or the test-bed environment, but there is often a huge gap in translating that research into products, services, best practices, and guidelines. Use-inspired research, combined with practitioners' willingness to test research results and implement effective strategies from the research lab, will benefit all of us involved in the challenges and rewards of digital preservation research.

References

All URLs were valid as of July 10, 2002.

National Science Foundation. 2002. *Revolutionizing Science and Technology through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure*, Draft 1.0 (April 19). Available at: http://worktools.si.umich.edu/workspaces/datkins/001.nsf.

Kenney, Anne R., and Oya Y. Reiger. 2000. *Moving Theory Into Practice*. Mountain View, Calif.: Research Libraries Group.

Stokes, Donald E. 1997. *Pasteur's Quadrant*. Washington, D.C.: Brookings Institution Press.

Understanding Digital Preservation: A Report from OCLC

Meg Bellinger

Overview

This paper reports on four aspects of the Online Computer Library Center's (OCLC's) current activities in digital preservation. Section 1 discusses recent strategic integration at OCLC to support digital preservation initiatives. Section 2 describes digital preservation activities of the Digital and Preservation Resources (DPR) centers, which are creating digital masters for the library community. Section 3 outlines technical considerations associated with building a Digital Archive, and Section 4 provides a list of activities in which the OCLC plans to engage with the digital preservation community.

Section 1: Strategic Integration

In September 2000, OCLC's Board approved a strategic plan under which libraries and OCLC will transform WorldCat from a bibliographic database and online union catalog into a globally networked information resource of text, graphics, sound, and motion. The rebirth of WorldCat in Oracle will create a global knowledge base supported by a set of integrated, Web-based tools and services that facilitate contribution, description, discovery, access, exchange, delivery, and preservation of knowledge objects as well as the expertise of participating institutions.

To realize this strategy, three of OCLC's primary business units are developing new products and services while enhancing current offerings. These units are Metadata and Cataloging Services, Cooperative Discovery Services, and DPR.

Acknowledgments
The author wishes to thank Taylor Surface, Pam Kircher, Leah Houser, and Linda Evers for their contributions.

Formalized in November 2001, DPR is OCLC's newest division and the topic of this paper. Today, 35 years after OCLC founder Frederick Kilgour's vision of pooling library resources began to be realized, DPR has taken on the task of building on his model. Our vision is to extend the OCLC cooperative to support the challenges of creating and sustaining access to and preservation of the global knowledge base's contents.

At present, DPR is home to three major initiatives:

- expanding Preservation Resources, a state-of-the-art preservation reformatting facility in Bethlehem, Pennsylvania, into regional DPR centers
- building a Digital Archive
- launching and supporting growth of the DPR Cooperative

This report focuses on how OCLC is expanding Preservation Resources' capabilities into regional DPR centers and construction of the Digital Archive.

Digital preservation takes place within a continuum, ideally starting from the point of digital-object creation, at a DPR center or elsewhere, and continuing through the processes involved in the long-term retention of those objects. DPR centers and the Digital Archive are two segments of the digital preservation continuum that have begun to converge as a result of OCLC's new business direction. That continuum is supported by the integration of infrastructure, metadata, and processes. If any of these three elements does not extend from one entity to another, preservation is not possible. The assumption is that a distributed, interoperable environment is the only viable approach to digital preservation. Digital preservation activities will occur both in the DPR centers and in the Digital Archive to support and reinforce the concept of preservation as a continuum of interdependent activities.

Section 2: Digital Mastering

With the creation of regional DPR centers, we are building on Preservation Resources' 15-plus years of experience with preservation microfilming. That translates into harnessing technology, skills, and processes for library-specific applications by creating a cost-effective, high-quality "digital factory" geared to meet the cultural heritage community's needs for digital preservation reformatting. We will maintain a test bed environment to experiment with digital imaging and metadata application processes in order to identify best practices, build tool sets, and anticipate future needs.

Recognizing the unique nature of materials in the information-services and cultural-heritage communities, Preservation Resources adapted commercially available technology to meet the needs of these communities. Adaptation in this case is expensive and somewhat difficult because the need is quite specific and, compared with that of the imaging industry as a whole, relatively small. As Kenney and Rieger state, "Determining how to digitize and present library

materials involves a fairly complex decision-making process that takes into consideration a range of issues, beginning with the nature of the source document but encompassing user needs, institutional goals and resources, and technological capabilities. These all map together as a matrix for making informed decisions rather than exacting standards" (2000, 24).

Preservation Digitization

Preservation Resources and DPR strive to support preservation librarians as they work through this complex decision-making process. The DPR centers are also developing the infrastructure with which to support the following assertion from the "Benchmark for Digital Reproductions of Monographs and Serials," endorsed by the Digital Library Foundation (2002): "Digital masters are digital objects that are optimally formatted and described with a view to their quality (functionality and use value), persistence (long-term access), and interoperability (e.g., across platforms and software environments)."

Complying with this benchmark requires a scanning environment capable of creating an accurate digital representation of our printed heritage. Consequently, we have determined that DPR centers must have three capacities. First, they must have material-handling skills to recognize bibliographic anomalies and other cultural representations. Second, they must have the technology with which to meet or exceed accepted quality standards. Finally, they must be able to engage in cost-effective standard setting for the broader commercial-service community.

Our challenge is to support this sophisticated scanning environment with technology and personnel to produce comparable levels of quality for various cultures, languages, and types of materials in DPR centers worldwide.

Preservation Metadata

Preservation metadata are any metadata used by an institution that is carrying out some form of digital preservation. Preservation metadata could include discovery, administrative, and structural metadata. Structural metadata should be sufficiently detailed to allow reconstruction of the sequence of the original artifact, a point that is being addressed as an addendum to the DLF benchmark (2002). More commonly, though, the term *preservation metadata* is applied to metadata serving either of two functions:

1. enabling preservation managers to take appropriate action to preserve a digital object's bit stream over the long term; or
2. ensuring that the content of the archived object can be rendered and interpreted.

Integrating digital preservation activities into the larger digital information life cycle and its associated workflows depends heavily on creating preservation metadata early in the process. Digital masters for digitally reformatted monographs and serials must have descriptive, structural, and administrative metadata, and the metadata

must be made available in well-documented formats. OCLC is likely to adopt the Metadata Encoding and Transmission Standard (METS) and create tools to apply METS in DPR centers at the point of digital-object creation.

To that end, staff at the DPR Center in Bethlehem, Pennsylvania, have created programs to automate population of the TIFF header with preservation metadata and are reviewing the NISO standards for still images. The issue of preservation metadata was addressed earlier in defining the Digital Archive system architecture (see Section 3 of this paper), but further work is clearly needed (see Section 4).

Authentication

Our community has much work ahead to develop processes at the point of digital-object creation that will support persistence. One area we are investigating is the high-volume, low-cost application of an authentication process at the point of creation. We may define *authentication* as a means for ascertaining that the digital material is what it purports to be and has not been altered since its creation.

Without data security, preservation is compromised. However, with the powerful flexibility of digital formats comes the ability to alter the original with ease and without detection. We are considering how to cost-effectively implement an authentication mechanism and are engaged in discussions about how to license and adapt third-party authentication software to our community's requirements.

The software we are evaluating functions basically as a digital notary public. The creator of a digital object uses the software to add a digital signature and time stamp to the object. That information is sent to the authentication software company for long-term retention. Future users can verify the security of the digital object by sending the registration information to the software company, which will determine whether it matches the original signature and time stamp on file. This service also records changes of ownership, further verifying its authenticity and providing a means of digital provenance. We will begin conducting an authentication pilot project with two library partners in the summer of 2002.

Section 3: Building a Digital Archive

A logical extension of various OCLC and Preservation Resources services is the construction of a Digital Archive. This activity focuses OCLC's longstanding strengths in research, software development, and cooperative work on the preservation mission.

Preservation Resources staff and computer scientists in OCLC's Office of Research have been working together since 1995 to understand better what is required for long-term digital preservation, both from a user perspective and a scalable, maintainable systems perspective. The current project to build the Digital Archive's infrastructure, metadata, and processes began in January 2001 and will proceed in multiple phases, as shown in figure 1.

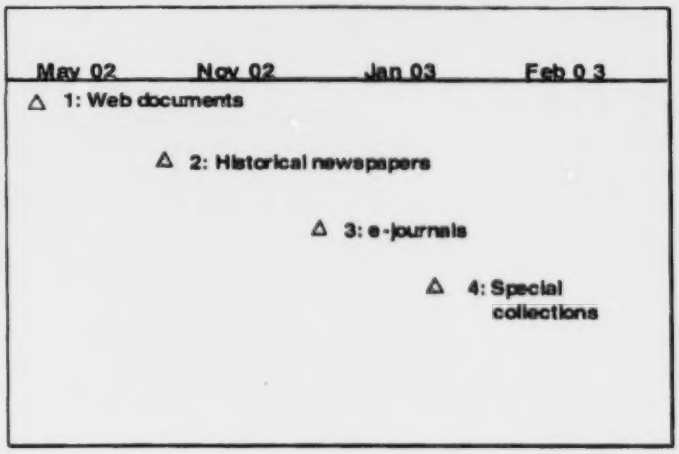


Fig. 1. The Digital Archive is being built in multiple phases

The project has three major goals:

1. To build a general-purpose digital archive for libraries, archives, and museums that may be used to store a variety of types of information and upon which various products and services may be built
2. To identify workflows for capturing and managing digital objects; and
3. To implement a metadata set for the archived objects

When Phase 1 is completed in May 2002, the system will facilitate the capture of Web documents, creation of preservation metadata for digital objects, ingestion of objects into the Digital Archive, and long-term retention of these digital information assets. However, this phase is limited in object format to text and still images. It is limited to ingesting objects into the archive one at a time, but it does have a set of tools that enable users to manage a complex workflow involving selection, cataloging, and archiving. The user can also generate a copy of the metadata and the object for in-house storage and dissemination.

Viewers will see objects in the Digital Archive by clicking on a URL in a bibliographic record in WorldCat, which they will access through FirstSearch, CORC, or a local catalog. They will also be able to access the Digital Archive by typing its URL into a Web browser.

Object owners will control access to their objects by creating content groups and related authorization groups. They will be able to delete their objects from the archive as well. For users familiar with OCLC's CORC, and FirstSearch interfaces, the system will be easy to use; however, the harvest software, the archive-object viewing interface, and the administration module have new interfaces.

Our decision to focus initially on Web documents was influenced by earlier work with the U.S. Government Printing Office (GPO) on a digital project. Having expressed the need to improve capture of Web-based government documents for long-term retention, the GPO was willing to work with us to define high-level user requirements for this data format.

As the project has progressed, we have involved other interested parties, mostly state libraries whose needs are similar to those of the GPO. Since 2001, the GPO has been joined by Ohio's Joint Electronic Records Repository Initiative, which includes the State Library of Ohio, the Ohio Historical Society, the Ohio Supercomputing Center, and the State of Ohio Department of Administrative Services; the Connecticut State Library; the Library of Michigan; Arizona State Library, Archives, and Public Records; and the University of Edinburgh, Scotland. Staff members from these institutions have met with us, commented on prototypes and workflows, provided input regarding the metadata element set, and participated in interface usability testing.

Preservation Metadata for the Digital Archive

Characteristics of objects and user groups are major factors in metadata decisions and in the tools created to support the metadata-creation process. The first objects in the OCLC Digital Archive will be born-digital and mostly public-domain government documents published on the Web and consisting of text and still images presented in HTML, PDF, JPEG, GIF, BMP, TIFF, and ASCII text formats.

Phase 1 users are mainly viewing objects created by others. As a result, they may not know of or not be able to obtain preservation metadata elements such as the recommended hardware for rendering an object. Also, our users want to integrate workflows to select, capture, catalog, and archive in a streamlined fashion. Finally, users want this integrated workflow to be as seamless as possible so that current staff can ingest objects and their preservation metadata into the archive efficiently.

Consequently, we have created new tools to make metadata creation easier, using as our foundation CORC, OCLC's tool set for creating descriptive metadata for electronic objects. CORC now supports a preservation metadata record that can be populated with data from a bibliographic record and updated with preservation data extracted from objects by the archive. Users may also enter data manually. We have also created a new harvester that launches from CORC and that uses tools within Oracle9*i* FS to extract technical information about the object. Finally, we are building a management module to enable users to assign objects to content groups and then specify access to that group.

OCLC staff kept these factors in mind when determining what preservation metadata elements would be needed in the first phase of the Digital Archive. These elements are as follows:

- user requirements
- object types
- tools

Some of the questions we asked ourselves were:

- What metadata are needed for these object types? (i.e., Web documents)
- When are the metadata available, and to whom?

- How are metadata captured, extracted, or created? By people or by a machine?
- How are the objects going to be accessed and by whom?

In answering those questions, we sought a balance among three elements:

1. preservation and maintenance of access to an object
2. what users can create practically
3. what the archive can extract or create

The Digital Archive's preservation metadata set is being developed by an OCLC team whose work is informed by the OCLC/RLG Working Group on Preservation Metadata as well as by other digital preservation initiatives. A report from the working group recommends a preservation metadata set and is available for review and comment (OCLC 2002).

When we compared the OCLC preservation metadata set with other element sets such as CEDARS or METS, we found that the convergences and issues for discussion were similar to the findings reported by the OCLC/RLG Working Group in its first white paper. To summarize those findings:

Convergences

- based on the Open Archival Information System (OAIS) reference model
- prescribes metadata for preservation
- able to extend the use of the archive to other object types

Issues for Discussion

- Scope: We are dealing with born-digital Web documents—other projects are dealing with converted materials or other formats.
- Granularity: We must determine at what level the metadata need to be assigned—logical object or file or both.
- Interoperability: This is an open question, but we are using an XML wrapper; communication with other groups is key.
- Implementation: While differences in implementation may not change how well the object is preserved, they may drive what tools are created for an archiving workflow or for accessing an object.

The OAIS Reference Model

Critical to enabling interoperability with other digital archives is our compliance with the OAIS reference model, which merits a brief explanation here. The International Organization for Standardization will soon publish the OAIS standard as ISO 14721:2002 (Garrett 2002).

OAIS grew out of the need of NASA and other national space agencies to capture, access, and store vast quantities of digital information for the long term. While the details of the reference model are complex, the overall concept is a straightforward sequence of input-process-output.

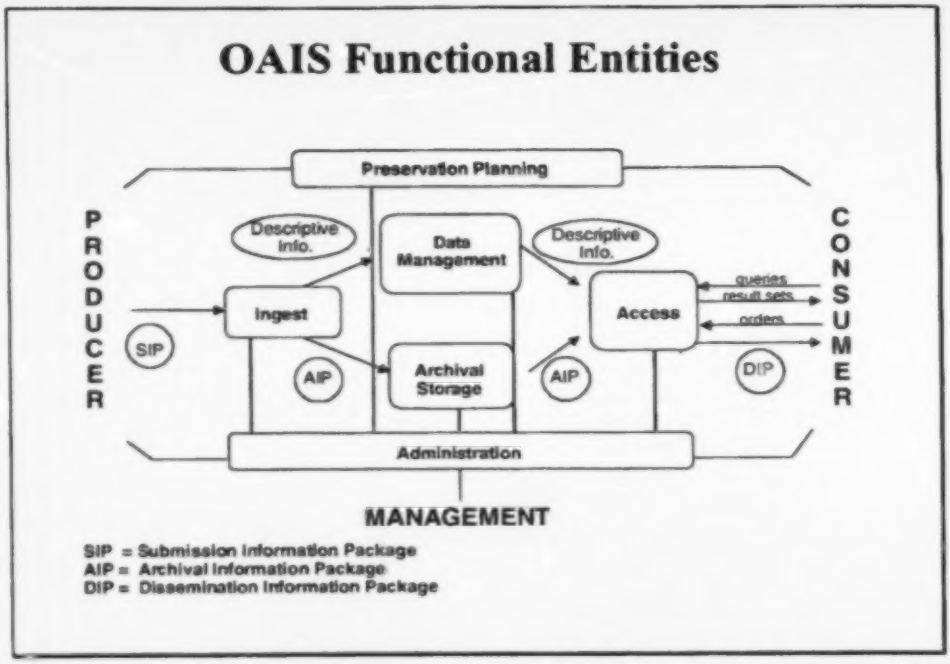


Fig. 2. The OAIS model has six functional areas and three types of information packages

Figure 2 depicts the conceptual relationships of the six functional areas and the three variations of information packages (Sawyer 2002). The sequence proceeds as follows:

1. A producer provides a submission information package (SIP) to the Ingest entity.
2. An archival information package (AIP) is created and delivered to Archival Storage.
3. Related descriptive information is provided to Data Management.
4. A consumer searches for and requests information using appropriate descriptive information and access aids.
5. The appropriate AIP is retrieved from Archival Storage and transformed by the Access entity into the appropriate dissemination information package (DIP) for delivery to the consumer.
6. Activities are carried out under the guidance of the Administration entity.
7. Preservation strategies and techniques are recommended by Preservation Planning and put in place by the Administration entity.

The three types of information packages are also shown in Figure 2:

1. A producer submits a SIP to the OAIS.
2. The OAIS holds and preserves the information using AIPs.
3. In response to consumer queries and resulting orders, DIPs are returned.

It may be useful to describe how we are implementing an OAIS-compliant AIP in the Digital Archive. Here is a list of the components we incorporated in the system architecture for AIP:

- 100 percent Java application developed by OCLC using Java beans, Enterprise JavaBeans, and JavaServer pages
- IBM AIX (UNIX) servers
- Oracle9*i* software

Oracle's content management products provide us with middleware, which we are using as a foundation for the Digital Archive. This middleware provides a Java abstract interface to the data repository for insertion, deletion, and other manipulation of content objects and metadata, including

- Object-level rights via ACL and ACE lists
- Fully extensible object-oriented database schema
- XML-enabled loading tools
- Extraction of structural metadata
- Metadata and future full-text searching
- HTTP, NFS, FTP, and Windows Explorer (SMB) protocol agents

The Digital Archive infrastructure builds on OCLC's existing procedures, staff, and environmentally controlled computer rooms. OCLC's experience in media migration keeps the bits alive as technology changes. Further, our experience with record conversions will be of good use in keeping both the metadata and content viable. As with WorldCat and other OCLC databases, copies of the Digital Archive's content and metadata will be stored securely in underground facilities off-site.

Every digital archive must plan for when vendors no longer support the tools with which it was built and is maintained and for when technological advances require a new system architecture. Consequently, our Digital Archive itself, like the data it holds, must have a planned migration path.

Under our plan for DPR's Digital Archive, we will extract objects and metadata from the repository in a system-neutral format to allow reloading into a new architecture. For significant system upgrades, support staff may be required to move data around, thus demonstrating their ability to do so in the event of completely new system architecture.

Section 4: Next Steps

As indicated in figure 1, we divided construction of the Digital Archive into multiple phases. Phases 2 and 3 will address the challenges posed by historical newspapers and e-journals; Phase 4 will probably focus on flat-text and still images. We have not yet identified the sequence in which we will accommodate audio, video, and dynamic formats such as relational databases and interactive instructional materials.

Inherent in each phase will be subsets of activities in which we intend to engage with the digital preservation community. Among these activities are investigation and development in several areas, including criteria for selection and required preservation-service lev-

el, new opportunities for cooperative activity in digital preservation, economic sustainability for digital repositories with preservation responsibilities, digital-rights management, preservation strategies, metadata requirements, and standards work based on the OAIS Reference Model.

Conclusion

Current trends in information technology and the emerging capabilities with which to build a global knowledge base offer exciting opportunities for libraries. Toward this end, DPR and other OCLC divisions are creating the tools and services libraries need to provide economical preservation of and access to materials. The expansion of Preservation Resources into DPR centers around the world and the construction of a large-scale, OAIS-compliant Digital Archive are tangible evidence of that work.

The magnitude of this task exceeds matters of hardware and software. We are aware of the need to build collaborative channels through which our members and the broader community can conveniently inform and immediately benefit from our ongoing work. Toward that end, we have launched and will continue to sponsor the DPR Cooperative, a group of diverse organizations and individuals who have joined us in accepting the challenge of exploring new opportunities in digital preservation.

OCLC takes as a profound responsibility the need that libraries and other organizations have to preserve cultural memory. Thus, it is imperative that we demonstrate the ability to provide a sustainable approach to long-term digital preservation and a commitment to do so with and for our community. This paper has described a methodology for expanding existing offerings and building new ones under a proven cost-recovery model. While these offerings will undergo transformations, we are building them with the belief that users in centuries to come will find our early collaborative efforts in digital preservation to have been worthwhile.

References

All URLs were valid as of July 10, 2002.

Digital Library Federation. 2002. Benchmark for Digital Reproductions of Monographs and Serials as Endorsed by the DLF. Available at http://www.diglib.org/standards/bmarkfin.htm.

Garrett, John (Web page curator). 2002. ISO Archiving Standards—Reference Model Papers. Available at http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.

Kenney, Anne R., and Oya Y. Rieger, editors and principal authors. 2000. *Moving Theory into Practice: Digital Imaging for Libraries and Archives.* Mountain View, Calif.: Research Libraries Group.

OCLC. April 2002. A Recommendation for Preservation Description Information: A Report by the OCLC/RLG Working Group on Preservation Metadata. Available at http://www.oclc.org/research/pmwg/pres_desc_info.pdf.

Sawyer, Donald M. 2002. Framework for Digital Archiving: OAIS Reference Model. Presentation delivered at the OCLC Steering by Standards Teleconference on the OAIS Imperative: Enduring Record or Digital Dust? Columbus, Ohio, April 19, 2002.

Update on the National Digital Infrastructure Initiative

Laura Campbell

For more than a decade, the Library of Congress (LC) has been trying to build a digital library. We now have more than 7.5 million items of historical significance online and available at no charge to schools and the general public. These materials represent a rich archive of American history, culture, and creativity. In addition, the LC has recently added material from five other countries. We intend to expand this program greatly. We have more than 102 archival collections online, and we will put another 20 collections online in 2002. These collections contain all kinds of new material.

In late 1998, the LC was thinking about the future. Were we ready to be in the digital business and to serve customers with new and different services? As part of that process, we brought together program managers from across the Library, which consists of the Copyright Office, the Congressional Research Service, the National Library Service Program, and the Law Library. We asked questions about how those managers saw their business changing. What kinds of programs and services did they envision for their customers in the future? Was the Library ready to stand up to the digital challenge of the twenty-first century?

Out of that process came a five-year plan that we presented to Congress in fiscal year 2000. The plan asked for $21.3 million to extend content, to enhance our infrastructure—our technology backbone—to build the components of a repository to house this material, to increase online access services (including digital reference services), and to continue our work with teachers online. Congress gave us about a third of what we asked for, but said, "Come back next year, and we will try to make you whole." We are now three years into that five-year plan, and the Congress has honored its commitment to provide what we originally requested.

Our focus has been sharpened and our efforts strengthened by the results of a report that had been commissioned in 1998 by Dr. James H. Billington, the Librarian of Congress. For that study, the National Research Council and the National Academy of Sciences assembled experts in the archival and information technology communities and the scholarly community to investigate whether the Library was well positioned to take on the digital task in the twenty-first century. That report, delivered in August 2000, was constructively critical about things we needed to pay attention to and the kinds of programs and activities we needed to put into place to learn from the broader community: digital content creators, distributors, and users—stakeholders with whom we had not traditionally dealt.

In the fall of 2000, representatives of the LC, along with Ken Thibodeau from the National Archives and Records Administration (NARA), presented a briefing to Congress. Supported by our five-year plan, the results of the National Research Council study, and our partnership with NARA, the Library was able to convince the Congress to authorize a $100-million special appropriation to be used to collaborate with the Department of Commerce, the National Archives, the White House Office of Science and Technology Policy, the other two national libraries (Agriculture and Medicine), the content community, the archival community, and the research community, as well as the technology community, to create a national strategy to collect, archive, and preserve digital content.

That legislation provides $5 million immediately for creating a master plan and $20 million contingent on approval of another plan that is targeted for submission late in 2002 to five congressional committees. As much as $75 million more may become available if it is matched, dollar for dollar, from non-federal sources. Thus, if the Library raises $75 million from private sources, Congress will also provide $75 million, bringing the total to $175 million. This match may be in-kind contributions or donations of services as well as cash. We believe that the matching component is most likely going to come from partnerships developed in the next phase of this effort to test potential models and options for long-term preservation.

The Big Picture: A Three-Phase Plan

Our plan has three phases. They are

- a preliminary phase, just described, resulting in a master plan to request Congress to approve the release of funds for investment in the broader community to test various approaches to the national strategy
- development of partnerships with the archival community and the content distributor/creator community
- a major effort to test and evaluate those partnerships and models that will enable the LC to go back to Congress in five to seven years to talk about the most sustainable options for long-term preservation

Our preliminary steps have included establishing a 26-member National Digital Library Advisory Board to which Council on Library and Information Resources (CLIR) President Deanna Marcum is a consultant. CLIR is helping us facilitate and work with the board, which is made up of many of the people cited in the legislation as well as experts in technology and publishing and people from the creator and distributor communities.

The board members have helped us develop ways to learn from this diverse stakeholder community and talk about the barriers to long-term preservation from their perspective. We have done a number of things to bring this group together, including coordinating with other federal agencies on a national research initiative. We have also undertaken systematic surveys and have examined other technical repository and preservation efforts, both domestic and international.

To organize our stakeholder community, we commissioned with CLIR six "environmental scans" of digital video, television, music, the Web, e-journals, and e-books. The results are available in print and on the Web (CLIR and LC 2002). Those scans, as well as about 20 confidential interviews with key members of various industry groups, helped us prepare for "convening sessions" conducted in the fall of 2001 in Washington, D.C. We held three two-day workshops that were identical in format, although very different in terms of discussion, to ask basic questions about barriers to the creation of a national strategy for long-term preservation.

These sessions helped us set priorities. Participants agreed about the need for a national preservation strategy. People from industry were receptive to the idea that the public good, as well as their own interests, would be served by coming together to think about long-term preservation. They also agreed on the need for some form of distributor-decentralized solution. Like others, they realize that no library can tackle the digital preservation challenge alone. Many parties will need to come together. Participants agreed about the need for digital preservation research, a clearer agenda, a better focus, and a greater appreciation that technology is not necessarily the prime focus. The big challenge might be organizational architecture, i.e., roles and responsibilities. Who is going to do what? How will we reach agreement?

Intellectual Property Remains a Concern

Other priorities were intellectual property and digital rights management. Defining the scope is likewise a concern: What is to be preserved, by whom, and at what level?

There are those who argue for "dark" archives; others are in favor of completely open access. Obviously, we have to strike a balance between preservation and access.

There are no widely used economic models for sustaining long-term preservation. Who is going to pay? Is preservation solely a government responsibility? Are there other ways to think about the economics of long-term preservation? Who will be the users?

We took the information from those sessions and created an agenda for what we call "scenario planning and analysis." This activity will be conducted by people who are skilled in looking at the future and the forces that may affect preservation, including such great uncertainties as government regulation.

Working with representatives of the Global Business Network, we created an agenda to bring before yet another group of industry experts. We talked about a timeframe of roughly 10 years. (It did not seem useful to go farther because we are struggling with what even the next three to five years might look like.) We talked about three possible scenarios: a "universal" library that collects everything, a library that is more selective, and a highly selective "world's-best" library.

Our planning with industry representatives has created a sense of urgency. We call it the "just do it" approach; its aim is to start collecting things before they are lost forever.

There was also a great emphasis on the need for distributed network technical architectures. In our first scenario-planning workshop, held in February 2002, we assembled a small group of technical experts who developed a hypothetical layered technical architecture. In the next step, we will think about how to fit into this architecture the functions and services that would be performed on a national level.

What Will People Take on Next?

We have tried to create a shared responsibility among the communities. What are people willing to take on next in planning? One of the initiatives has great momentum. Margaret Hedstrom, from the University of Michigan, Donald Waters, from The Andrew W. Mellon Foundation, and a number of other people have made it possible for us to bring together, working with the National Science Foundation, some 50 experts from 15 federal agencies and the private research laboratory, technology, and computer science communities. This group is shaping a focused agenda on digital preservation, leveraging our collective resources, and bringing together funding from various agencies to put digital preservation research within a framework that can serve many of us, not just some of us.

That effort led to a number of informal meetings and to a formal workshop that was held in Washington, D.C., in April 2002. It was a good beginning: the executive branch of the government was represented by individuals from the Central Intelligence Agency and the Department of Defense; the LC represented the legislative branch. Also attending were representatives from other national libraries, the scientific community, and the library community. Margaret Hedstrom is the principal investigator. We will end up with a call for proposals this fall and, we hope, provide funding to the best initiatives.

As part of our early planning, we have developed a conceptual framework to think about the components—the political, economic, social, legal, technical, and organizational concerns—that need to be considered in a national strategy. Information about this framework

is available on our Web site, referenced below. We have also tried to identify a set of critical technical issues on the basis of what we have heard in the surveys and our benchmarking. A summary of this work is also available on our Web site.

Later this year, we hope to incorporate what we have learned from the industries—the roles they are willing to play and issues on which we should be focusing—into our national strategy for preservation. This strategy will include credible scenarios and models based on our sense of who is willing to do what, with whom, and for what purpose. Our goal is to make progress toward a national preservation strategy and a research agenda that are grounded in an investment framework that Congress will understand.

The process will yield a "master plan." Once the plan has been developed, we will present it to five congressional committees. Our presentation will also include expert testimony. We trust these efforts demonstrate that the LC is being responsible—bringing the right parties together to recommend the areas in which we should invest to meet the national need to preserve material for the future.

References

All URLs were valid as of July 10, 2002.

Council on Library and Information Resources and Library of Congress. 2002. *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington, D.C.: Council on Library and Information Resources and Library of Congress. Also available on the Digital Preservation Web site, http://www.digitalpreservation.gov.

Experience of the National Library of the Netherlands

Titia van der Werf

Understanding of the issues surrounding digital preservation is growing at a reassuring rate, at least among the nucleus of experts in this field. At the same time, confusion and a sense of being overwhelmed by the complexity of the issues often prevails in the wider community of archives and libraries. The national approaches that are now being started in the United Kingdom, such as the Digital Preservation Coalition initiative, as well as in countries with national digital strategies, such as the United States, are commendable; at the same time, they can distract institutions from just going ahead and acting. Moreover, some organizations, national archives as well as national libraries, seem to be stuck in the requirements-specification stage and find it difficult to move forward to implementation, perhaps out of fear of making mistakes.

In the Netherlands, we do not have a national strategy yet, but we have advanced quite a bit at the institutional level, especially in my library. The archive community, together with the Ministry of Home Affairs, has been setting up test beds for digital preservation. This paper focuses on two activities—how we are preparing to create a mass storage system as well as the work we are doing with IBM on long-term digital preservation issues. Like others, we have made mistakes, but we have also made substantial progress. This paper describes what we have done and the lessons learned en route.

Serious Business: Preparing a Mass Storage System

In contrast to most other countries, the Netherlands does not have a legal deposit regime. The national library of the Netherlands makes voluntary deposit agreements with the Dutch Publishers Association. Our agreements with publishers date from about 1974 for printed publications and from 1995 for electronic publications. The latter

include offline as well as online publications. To date, we have collected monographs, CD-ROM titles, government reports, and dissertations. Among the serial collections are online journal articles from Elsevier Science and Kluwer Academic Publishers, and we have also received official government publications in digital form.

Altogether, our collection amounts to three terabytes of information. All the CD-ROM information is still in the offline medium; we have not yet transferred it to a storage system. We do not have enough storage capacity in our digital stacks for all digital deposit material. Because we anticipate a great deal of growth and the receipt of all the Elsevier titles in the coming year, getting a reliable, quality-controlled mass storage system in place is one of our priorities.

Different Countries, Common Goals

One of our most important activities has been leading the Networked European Deposit Library (NEDLIB) Project. As this project got under way, the eight participating national libraries entered into discussions about how to set up a digital deposit system. We spent a year talking about our differences. While time-consuming, these discussions were necessary because they gave us a common understanding of issues at stake. We slowly realized that we needed to identify our common missions, goals, and objectives. We asked "What is common to the digital publications that we receive, and what common solutions can we come up with for this problem?" This exercise laid the foundation for the consensus building that would occur later in the project.

We looked at the deposit process, that is, the workflow for electronic publications. First, a publication gets selected; then it comes in as a deposit and we capture, describe, and identify it. This process continues through the whole workflow, ending at user services. Next, we identified areas where we thought that this process might be different in the digital world than it is in the world of print material. For example, in the digital world, our library has no system in place that could tackle the new parts of the process required by digital publications. We have automated cataloging systems and acquisition systems, but even if we use these systems for digital publications as well, we still do not have a storage system or a digital preservation system.

We identified the missing information technology (IT) components. Figure 1 shows how we visualize the existing library systems that support the conventional steps of the workflow. For the steps in this workflow that are not supported by any current system, we realized that we would need to put a new system in place. By following the workflow steps, we could start identifying our requirements for this new system.

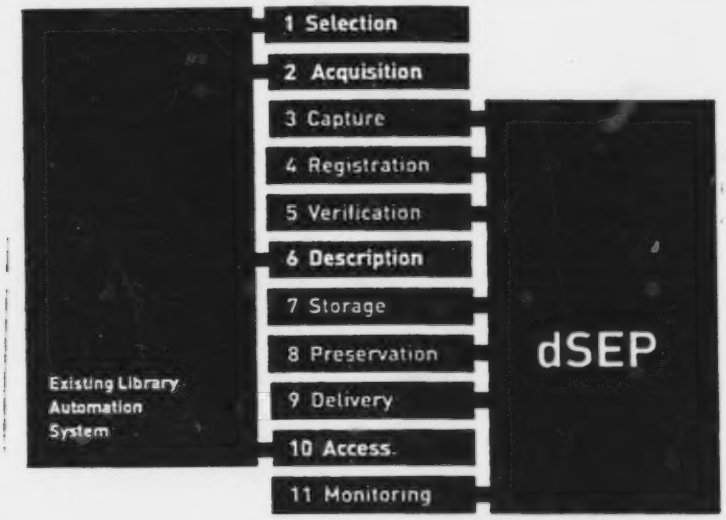


Fig. 1. The NEDLIB workflow for electronic publications

NEDLIB has now developed guidelines for national libraries that want to set up digital deposit collections and systems. The guidelines outline how to design the digital stacks as a separate module in the digital library infrastructure and how to implement the digital stacks as much as possible in conformity with the Open Archival Information System (OAIS) standard. We had been looking extensively at the OAIS standard in the NEDLIB project and were impressed that this model suited our requirements so well.

We also recognized that electronic publications coming in should be transferred to a highly controlled storage environment, and that after they are brought into the digital stacks, all objects should be treated alike. That led to another question: Should we implement separate systems for Web archiving, CD-ROM archiving, and archiving of online journals? We realized that if we did, the process could go on forever; we would have hundreds of systems sitting next to each other. That would not be manageable. Therefore, we decided that we wanted to be able to treat all electronic publications in the same manner, regardless of type.

When the NEDLIB project ended, we returned to our own countries and started implementing local digital stack systems in our libraries. We are all implementing these local deposit systems in different ways, but we hope that through NEDLIB we have gained a common understanding of what we are doing.

IBM's Implementation Model: A Work in Progress

In the Netherlands we issued a request for information that asked the IT market whether there were products that could provide for system functions according to the NEDLIB/OAIS Model. As a result of the positive reactions from the IT sector, we started a tendering procedure.

BEST COPY AVAILABLE

IBM Netherlands, which had off-the-shelf products that could support quite a few of the processes that we had identified, was the successful candidate. IBM made it clear from the start that its products would not be able to provide any long-term digital preservation functionality, but it was willing to help us research the issues and look at the requirements of this subsystem for preservation.[1]

Figure 2 shows the IBM Implementation Model. It depicts the OAIS modules of Ingest, Archival Storage, Access, Data Management and Administration, as well as the NEDLIB-added modules of Preservation, Delivery and Capture, and Packaging and Delivery. The latter two modules are interfacing with existing library systems. Within these interfacing modules are gathered everything that has to do with locally defined and variable types of things. For example, at the input side are all the different file formats that publishers use. Because these formats are not generic and change over time, we regard them as external variables instead of internal archive standards. At the output side are different types of customers. The requested archive objects must be tailored to make them fit for use, both now and in the future. Library users' groups will change over time, and users will become increasingly demanding as technology evolves. In summary, we put everything that is variable into these interfacing modules and everything that is generic into the "black box" that we call our deposit system.

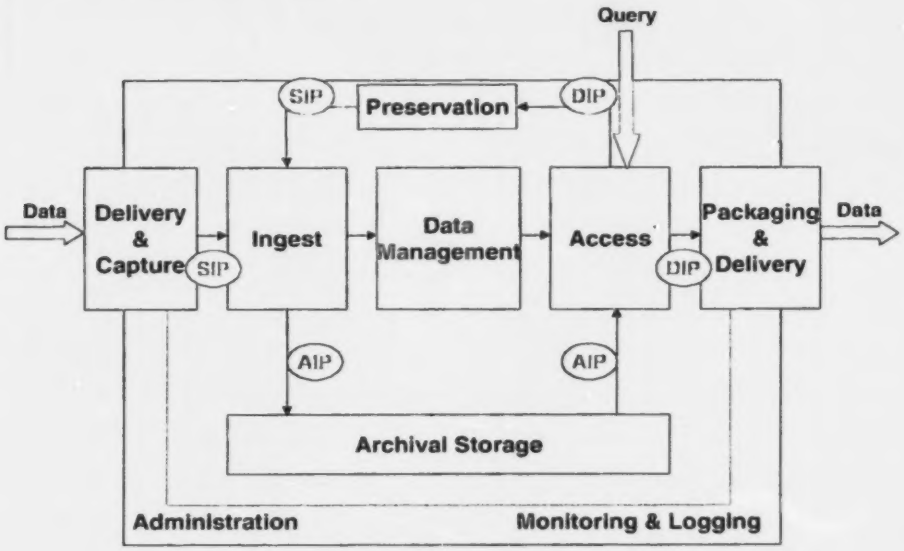


Fig. 2. IBM Implementation Model

IBM has named its system implementation the Digital Archival Information System (DAIS). We have DAIS Version 1.0, whose scope includes a pre-Ingest module for manual loading of submission information packages (SIPs). With this module we can load CD-ROMs and e-books. There is an archival storage unit with backup, disaster

[1] Additional information on the collaboration between the national library of the Netherlands and IBM-Netherlands can be found in Niewenburg 2001.

BEST COPY AVAILABLE

recovery facilities, and a data management unit where technical metadata are recorded. Descriptive metadata and structural metadata are put into other existing library systems. The Access module for Dissemination Information Package (DIP) retrieval is complemented with a post-retrieval module for installation of electronic publications on the library workstations.

We are already planning for DAIS Version 2.0 as we anticipate additional functional requirements. We know we need full automatic batch loading of SIPs, especially for the e-journals that come in with a certain regularity and frequency. We do not want to do that all by hand, article by article. With DAIS, serials processing and archiving will be automated.

In Version 1.0, we are able to add new SIPs to the system. We know from experience, however, that some submissions need to be replaced or even deleted, even though it is a deposit. For that reason, we also need some replace and delete functionality in the DAIS-system.

IBM will upgrade the system to Content Manager Version 8.0, thereby adding new functionality as well. The British Library, which is involved in a similar effort with IBM-UK, has expressed its intention to build its deposit system on top of DAIS Version 1.0. Finally, the preservation subsystem still needs to be implemented in future versions of DAIS.

Defining Scope to Avoid Distraction

The IBM implementation has raised many issues outside its immediate scope. Just because we have a system does not mean that it will support the whole workflow. DAIS is only one piece of a large puzzle, and the pieces we still need include, for example, a Uniform Resource Name (URN)-based object identifier system.

We also need batch processing of the associated metadata we receive from the publishers, together with the content files. The metadata should be processed automatically and converted to our own XML DTD format, and they should be loaded into our metadata repository and indexed.

We need a digital mailroom where publishers can deposit their electronic publications. This should be a controlled area with password, virus checking, mirroring, and harvesting capabilities. We also need a preload area for both batch and manual loading. Other needs include user identification and authentication, authorization mechanisms, and collection browsing functionality.

Although we need these functions, we ultimately did not ask IBM to include them in their implementation system. We realized that adding new functionality along the way would jeopardize the project budget and schedule. Also we wanted a modular architecture, with well-defined functions and supporting technologies. We investigated whether we could implement some things ourselves, or whether there were products on the market that would be suitable. This scoping effort is very important to make sure that you take only

the generic parts into the system and have a modular way of building your digital library infrastructure.

What are the digital stacks? Essentially, it is the IBM system and possibly other content management systems as well. We are thinking of implementing a separate system for our digitization collections because we are not going to put those collections in the IBM system.

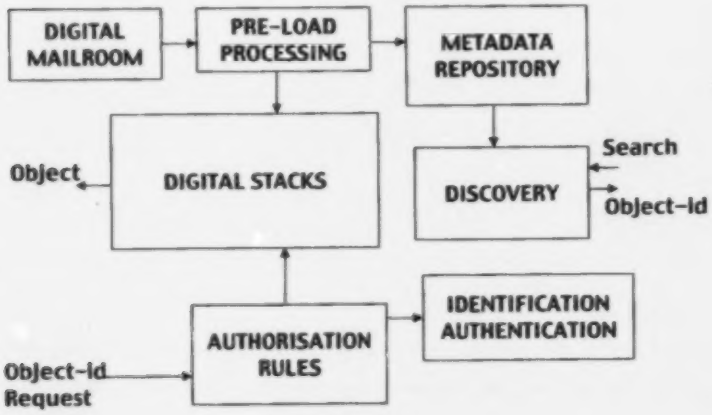


Fig. 3. Digital Library Infrastructure

Deciding whether to add our digitized collections to the IBM system has been difficult. Our primary goal was the deposit collection. The requirements, therefore, were for managing highly complex and controlled technical metadata to ensure that later on we would be able to migrate, convert, or emulate formats. We had developed several requirements relating to the pre-ingest and post-retrieval processes that were much too heavy for the digitized collections. Digitization is all about providing quick access and putting it quickly on the Web. This is in contrast to requirements for deposit collections. We ultimately decided to create a separate system for the digitized collections.

A Data Model for Long-Term Access

The data model implemented in the IBM system is based on the OAIS model. In one archival information package (AIP), we envision being able to put in either the original version of the electronic publication or a converted version of the electronic publication. Also, we envision keeping the bit-images of installed publications in AIPs. For example, if a CD-ROM publication needs to be installed, we take a clean workstation configuration and install the CD-ROM on it. Then we take a bootable image of the installed CD-ROM and enter that as one image in the AIP. This process allows for access now and, we hope, in the future.

We also envision putting software applications (including platform emulations and virtual machines) in AIPs and operating systems, if need be, as disk images. Of course, hardware platforms cannot be put in an AIP.

In DAIS Version 1.0 we have in place as many long-term preservation hooks as we could think of. The data model, and especially the technical metadata, are important hooks. In terms of technical metadata, we are concentrating on what we really need to record about the technical dependencies between file formats, software applications, operating systems, and hardware platforms. We see this as a prerequisite to ensure access and readability now and in the future.

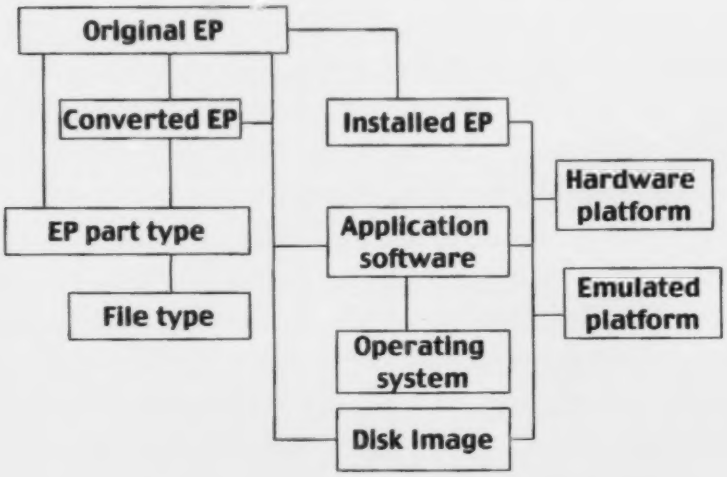


Fig. 4. Data model of the IBM implementation

A Reference Platform for Manageability

As Ken Thibodeau has said, when you start registering all this technical information and look at each file type for what system it can run on, you find that it can run on up to 100 different systems or configurations. Recording all that will take a good deal of time, and that is what he was visualizing when he created his graph with preservation methods and format types. He put preserving original technology at the "great variety" end of the continuum and contrasted it to preserving the more generic and persistent content of objects.

Recognizing this problem, we considered managing the technical metadata in terms of the concept of a "reference platform." The reference platform concept tries to freeze the configuration of a workstation or PC for a whole generation of electronic publications—perhaps for five years. The configuration includes the hardware, the operating system (e.g., Microsoft Windows 2000), viewer applications (e.g., an Acrobat reader), and a Web browser, as is shown in Figure 5.

This frozen workstation would cater to a generation of publications: for example, all PDFs published between 1998 and 2002. Everything we receive in the library currently is in PDF format. We hope that with this frozen workstation, we will be able to manage the diversity in configurations that may appear during this period of time. We do not want to support all possible view paths, just the pre-

ferred view path. The reference workstation is the preferred view path for a collection of publications for a certain period of time. In this way, we can standardize the view paths we support in our system and make the handling of diverse configurations more manageable.

The technical metadata records the chain of software and hardware dependencies. We create technical metadata to be able to recreate the running environment for access now and in the future. This process will also help us monitor technological obsolescence. The reference platform is a means to make all this more manageable.

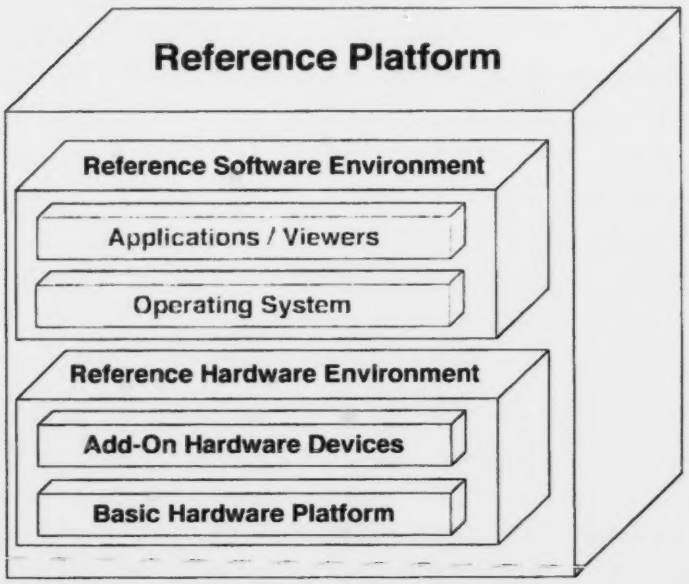


Fig. 5. The Concept of the Reference Platform

New Skills Call for New Jobs

With the workflow for electronic publications, we need new skills and staff. We need staff for the daily operation of the digital stacks, most likely not organized as a new department but as part of our existing processing department where the general cataloging takes place. We need technical catalogers who can handle many different file types and IT configuration management tools. We need technology watch officers who monitor for new formats and trends on the e-publishing market. Further, we need reference platform administrators, digital preservation researchers, and quality control managers.

Assistance from the Computer Scientists

The Koninklijke Bibliotheek has worked with Jeff Rothenberg of The RAND Corporation and IBM's Raymond Lorie. We began working with Jeff Rothenberg because his emulation theory was new to us and we were interested in solutions. When we asked him to explain his solution, he presented his hypothesis as it stood in 1995 (Rothenberg 1995). As he talked with us and with representatives of other

memory institutions, such as the Dutch National Archives, he developed his concept to maturity (Rothenberg 2000). He addressed the problem of needing to build different versions of emulators over time. Consequently, he developed the idea of the virtual machine on which you would be able to extend the life of an emulator.

With Raymond Lorie, we looked at all the PDF files in our digital deposit. He worked on a prototype to demonstrate that you could extract both the data and the logical structure of the document to recreate it at access time, even without the original PDF file. What you would need to keep is the bit-image of the document and a Universal Virtual Computer (UVC) program to interpret the logical data structure of this document, so that you would be able to scroll, search for words, and navigate through the document.

Lorie's approach offers a generic way of handling PDFs that could be applicable to other types of data formats, such as JPEG (Lorie 2001). With this approach, you could discard the original PDF, although we have not chosen to do so. We agree with Ken Thibodeau when he says that as long as we can afford it, we should try and keep everything in its original state, even if we know that we will not be able to read it in 10 years' time. Maybe in 150 years we will be able to read it; who knows?

Talking with computer specialists such as Jeff Rothenberg and Raymond Lorie helped us understand the need to distinguish between two types of electronic publications: executable publications and document-like publications. Executable publications are software programs; they include games and complex publications on CD-ROM. Document-like publications are the simpler data types, such as texts and images, which require only viewers to interpret them. Programs and document-like objects require different technology solutions. We have been investigating virtual machine technology and hardware emulation to solve the problem of hardware dependency, and we have been looking at data extraction for solving the software dependency problem.

Mistakes Bring Wisdom: Lessons Learned

What lessons have we learned in the course of the activities just described? First, the question of scope is important, because otherwise it is easy to get distracted. At one point we mentioned to IBM that our digitized collection needed to be put into this system. When they asked us about our requirements for that type of collection, we answered, "Fast access and fast ingest." Those requirements, however, contradicted those we had previously specified for our deposit collection. This was confusing, and it took much precious time to clear up the confusion. We decided to forget about the digitized collection and return to our original goal. This shows that you just cannot try to tackle all the problems at once. A step-by-step approach is essential. This does not mean, however, that you can overlook the need for a comprehensive approach, because you always need to keep the big picture in mind.

Modular design is also important, especially the ability to have independent modules so that you use the right technology for the right module and do not create unnecessary dependencies. Each technology—for cataloging, indexing, or storage, for example—changes at its own pace. Technologies should not be too dependent on each other, because when you change one technology, you will have to upgrade the whole system.

It is important to think about whether you are going to choose IT market solutions or develop your own. We have reorganized our IT department in such a way that we no longer support IT development. Instead, we outsource everything pertaining to development or hire people with the needed development skills. Our policy is to rely as much as possible on IT market products rather than custom-made products, so that we can gain leverage from IT product support and development services.

Technology Problems Need Technology Solutions

Analysis of a list of digital preservation projects recently drawn up by the European Union-National Science Foundation Working Group shows that not all these projects are about long-term digital preservation. Many are about building controlled archives, which is really much more about storage management, rather than long-term preservation. The "Lots of Copies Keep Stuff Safe" (LOCKSS) approach, for example, has often been cited as the answer for digital preservation. But it is not the answer, because it is really nothing more than a very controlled (or maybe uncontrolled!) way of replicating. LOCKSS does not preserve anything in the long term. If a format is obsolescent now, it will still be obsolescent in the future.

There is an important difference between archiving and long-term access. Archiving is quite straightforward because we are doing it already: selecting, identifying, describing, storing, and managing. Archiving keeps the objects machine-readable and healthy, and it provides access now. The real challenge is long-term access—being able to render the object in a human-understandable form and being able to solve the software and hardware dependencies.

Digital preservation is a technology problem, and it needs a technology solution. Metadata are a means to help us solve the problem, but digital preservation is not a metadata issue. It is the same with access control issues. People talk about dark, dim, and bright archives, but that has to do with access control, not long-term digital preservation.

Spreading the Word for a Shared Problem

Digital preservation is not only the problem of memory institutions. We have new players and potential partners in our midst, such as Warner Brothers, a film business partner. The problem we are tackling is a shared problem across our new, e-based society. Businesses, public service and health organizations, schools, and research insti-

tutions have a stake in this issue, as do individuals. You have your personal files, including electronic income tax files and Web pages, that you want to keep for a longer period of time—the tax files because you are required to keep them for at least five years, the digital pictures and Web pages for your grandchildren.

We, as memory organizations, are responsible for raising awareness. Where other sectors in society are not yet fully aware of the need to develop a digital memory, we are the ones that can raise this awareness. The adoption of a Resolution on Preserving our Digital Heritage at the UNESCO general conference in October 2001 has been a very important awareness-raising step.

How can you measure the state of digital preservation? What do we have in place to be able to measure at what development stage we have arrived? It might be useful to look for similarities in another field, such as medicine. The first question is "Do we have patients?" Yes, we have an increasing number of digital archives and collections that are potentially endangered. Do we have doctors? Yes, we have increasing numbers of experts. What are the illnesses? Do we know the symptoms? We do know that there are increasing examples of digital obsolescence, of tapes that cannot be read, file types that are no longer supported by any software, but we do not know how many. Are there research programs? Yes. Research is under way worldwide, and we know that much more needs to be done. We are at the stage of drawing up research agendas. Do we have hospitals? Data recovery centers? I think there are one or two in the world.

Do we know the treatment against the illnesses? Are there medicines and cures? We have advisors, we have best practices, but not much more than that. Educational programs? Yes, in rising numbers. Do we have emergency kits? These are all questions that beg for answers, and only in raising these questions can we raise awareness of digital preservation issues across society and across the globe.

References

All URLs were valid as of July 10, 2002.

Lorie, Raymond A. 2001. Long Term Preservation of Digital Information. In: *Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia; p.346-352. New York: ACM-Press.

Nieuwenburg, Betty. 2001. Solving Long Term Access for Electronic Publications. D-Lib Magazine 7(11). Available at http://www.dlib.org/dlib/november01/11inbrief.html#NIEUWENBURG.

Rothenberg, Jeff. 1995. Ensuring the Longevity of Digital Documents. *Scientific American*, 272(1): 42–7 (international edition, pp. 24–9).

Rothenberg, Jeff. 2000. An Experiment in Using Emulation to Preserve Digital Publications. RAND-Europe. NEDLIB Report Series;1. Den Haag: Koninklijke Bibliotheek.

Digital Preservation—A Many-Layered Thing: Experience at the National Library of Australia

Colin Webb

The National Library of Australia (NLA) has made a substantial contribution to digital preservation practice, research, and thinking. In 1995, it established one of the world's first library digital preservation sections. Our PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) archive of online publications, in operation since 1996, has developed into a collaborative national collection. The NLA was an early contributor to international discussion and debate through negotiation of a Statement of Principles for the Preservation of and Long-Term Access to Australian Digital Objects (NLA 1997). The Library also provided input to the Task Force on Archiving of Digital Information set up by the Commission on Preservation and Access and the Research Libraries Group (1996). The NLA has played a leading role in raising and discussing digital preservation issues with other national libraries and in looking for areas of collaboration (Fullerton 1998). Finally, the Library's Preserving Access to Digital Information (PADI) Web site is a tool for keeping up to date on digital preservation developments worldwide.

While these and many other initiatives are evidence of the NLA's commitment to promoting digital preservation practice and thinking, their main value in the context of this symposium resides in the lessons and principles that can be drawn from them. The purpose of this paper is to analyze experience at the National Library of Australia and to determine whether it can offer anything of value to other libraries' digital preservation activities.

Acknowledgments
The assistance of the Council on Library and Information Resources (CLIR), the Digital Library Federation, and Documentation Abstracts, Inc., in supporting my participation in this symposium is gratefully acknowledged. This paper draws in places on the work of colleagues at the National Library of Australia, including Margaret Phillips, Kevin Bradley, and Gerard Clifton.

Unpacking the Digital Preservation Problem

In Australia as elsewhere, library professionals have been engaged in a quest to find effective solutions to an overwhelmingly complex problem: the preservation of digital information. Experience at the NLA suggests that it is profitable to see digital preservation not as a monolithic problem but as a challenge with many layers. For us, approaching digital preservation from this perspective has been enormously productive.

One set of layers concerns *different types of digital collections*. For NLA, these collections include online publications, physical format digital publications, digital sound files, image files, corporate records, and a number of other discrete collections. While we recognize that all digital data can be handled the same way, it has taken us nearly 10 years even to start implementing systems that will integrate the management of these collections. It will take more years before we achieve full integration; perhaps we will never do so. This situation reflects our collection-oriented approach: for example, we organize publications collected from the Internet in a different manner than we do the oral history audio files we create; and the way we provide access to corporate records is different from the way we handle open-access collections. Our approach also reflects some technical differences; for example, the range of file formats and software dependencies in the material we collect is much more diverse than that of the digital materials we create ourselves.

For pragmatic reasons, we began by setting up what we believed would be the best way to manage each type of collection, knowing we would have opportunities to bring these systems together as we built knowledge and system capabilities over time.

A second set of layers concerns *stages of action*. We decided to begin by addressing immediate issues that threatened to rob us of the chance of taking longer-term preservation action. We made a conscious decision to take first steps—intelligent first steps if possible—without knowing all the challenges we would face or how to solve them.

Our most pressing demands were to make some decisions about what we should try to preserve and to put those materials in a safe place. We subsequently added layers of description, control, access, preservation planning, and action, and we are gradually integrating processes. However, the ability to look for staged responses and, when necessary, to separate processes, remains key to understanding and tackling problems.

We have formalized these processes into two broad terms: *archiving* and *long-term preservation*. For the NLA, *archiving* refers to the process of bringing material into an archive; *long-term preservation* refers to the process of ensuring that archived material remains authentic and accessible. Thus, we quite happily have a manager of digital archiving and a manager of digital preservation who work closely together and understand the subtle but important differences between their roles.

A third set of layers concerns *levels of action*. We found that we could and should distinguish among intentions, commitments, actions being planned, and actions proven and in place. The differences among these levels are easily and dangerously blurred; for example, when we assume material has been preserved simply because it has been saved to an archive. When we have recognized the differences among these levels, it has spurred us forward.

Infrastructure often seems to be the main enabler in moving from one action level to another. At NLA, we define *infrastructure* as the tools and systems for managing digital collections, the policies that guide what we do, and our means of sharing information, agreements, and accountability measures. Developing infrastructure takes time and resources, but it is a necessary investment. By developing infrastructure in parallel with archiving action, we have allowed for feedback processes so both activities can inform each other. It is not always necessary, or even desirable, to wait until all the infrastructure is in place before beginning the preservation activity.

A final set of layers concerns *responsibility*. We were impressed with the approach taken by the Library of the University of California at Berkeley and described in its 1996 Digital Library SunSITE Collection and Preservation Policy (Library of the University of California, Berkeley, 1996). This policy makes a clear distinction between the resources for which it will take archiving responsibility and the other materials available from its site. In trying to establish a distributed, collaborative approach to managing a collection of online publications, NLA has explored responsibility principles expressed in down-to-earth terms such as

- "Everyone doesn't have to do everything."
- "We don't have to do everything all at once."
- "Responsibility can be time limited: it doesn't have to be forever for everyone."

These principles will be addressed in greater detail in the section entitled "A National Model."

To summarize, the concept of *layers* underlies much of the National Library of Australia's progress in responding to the challenges of digital preservation. While we have often been characterized as advocates of a "just do it" approach, we believe that "just doing it" can be carried out in a systematic, intelligent, and learning-oriented way.

Situational Factors

The approach just described is one that tries to respond to the real world of digital information within a framework of evolving conceptualization. Under such an approach, it would be inconsistent to expect the experience in digital preservation at one institution to be a sure guide for every other program. The circumstances in which we operate function as constraints and enablers that help define what we want to achieve and how we go about it. Some explanation of the

NLA's circumstances may help others understand what we do and identify commonalities and differences with their own experience.

Australia is a large country, similar in size to the United States but with a much smaller population. Roughly 20 million Australians live in a number of large urban sprawls scattered around the fertile edges of the continent, in regional towns and cities, or in remote communities. Almost every political jurisdiction is characterized by a dichotomy between relatively large urban populations and what we call "the bush," whose inhabitants often have very limited access to the information resources accessible in the cities.

To some extent, the Australian library system reflects these geographic realities. The system also reflects our national history and the foundations of pre-existing colonies that carried many of their roles with them into Federation in 1901. Central libraries with deposit functions and public library systems committed to serve the populace wherever they could efficiently do so were a part of Australia's history. Libraries have a proud place in the Australian ideal of a fair, open, and educated society in which there is both equality of opportunity and reward for initiative and excellence. Such an idealized picture has often been undercut by realities, including the fact that many Australians are denied equality of access to information because of distance, income, education, or background.

In such an environment sits the NLA—working with, leading, and serving a library system made up of many autonomous parts geographically distant from the majority of Australians who own it through their taxes, and committed to providing effective information services to all Australians.

It is not surprising that Australians have taken up digital technology and that institutions such as the NLA see the exploitation of digital information as critical to their futures. Without embracing digital information, and without managing and preserving digital resources, the NLA would face increasing irrelevance. Thus, in the 1990s, the Library made a deliberate choice to bring digital information resources and services into its core business. From this decision flows virtually all progress the Library has made in building and managing digital collections and in working with others engaged in similar work.

The NLA is established by law and largely funded by annual federal government appropriations to deliver a number of functions including

- developing and maintaining a national collection relating to Australia and the Australian people
- making material from its collections available for use
- cooperating in library matters with others in Australia and elsewhere

Bringing the management of digital information resources into the Library's core business means applying these functions to such resources.

Digital Collections

The NLA manages a range of digital collections. While most attention has been paid to one of these—the PANDORA archive—our programs seek to manage all of the collections for which the Library accepts long-term responsibility. These collections include

- Online publications selected for the National Collection of Australian Online Publications managed in the PANDORA archive. Establishment and management of this collection are described in detail in the sections titled "Collection Building" and "Digital Preservation."
- Physical format digital publications (distributed on diskettes or CD-ROMs). Preservation actions for this collection are described under "Digital Preservation."
- Oral history sound recordings. Preservation actions for this collection are also described under "Digital Preservation."
- Both intentionally and unintentionally deposited manuscript materials on digital carriers. Recovery procedures for inaccessible items are briefly described under "Digital Preservation."
- Digital copies of analog collection items produced in our digitization programs.
- "Born-digital" unpublished pictorial works such as photographs.
- Corporate electronic records.
- Bibliographic and other metadata records.

Collection Building: PANDORA as an Example

Most of the NLA's collection-building activity has involved online publications. For this reason, the following discussion focuses on the PANDORA archive. The PANDORA archive of Australian online publications has been described in many papers available from the NLA Web site (Cathro 2001). This discussion is limited to the points that are most relevant to a broad understanding of what PANDORA is and how it works. Although initiated and managed by the NLA, PANDORA has in recent years developed into a collaboration among a number of partners.

The origins of PANDORA lie in the conviction that the Library has a responsibility to collect and preserve the published national heritage, regardless of format. The Library started discussing options for preserving online electronic information resources in the early 1990s. In spite of predictions that it would be technically too hard and that there would be insurmountable copyright obstacles, we decided to take some exploratory steps and see what progress we could make. Thus, in 1995-96 we appointed an electronic preservation specialist in our Preservation Services branch; set up a cross-program committee to develop guidelines for selecting online publications that should be collected; established an Electronic Unit to select and catalog online publications; and began to experiment with capturing (and sometimes losing) selected publications using cobbled-together, public domain software.

From these uncertain beginnings, PANDORA has developed into an operational National Collection of Australian Online Publications. It contains about 2,200 titles, roughly half of which have multiple instances (i.e., they have been gathered more than once). Roughly a third of the titles in the archive are collected on a regular basis; however, the frequency of capture varies, depending on the gathering regime negotiated with the publication owners.

To build a national collection, NLA works with a number of partners, including ScreenSound Australia (the national film and sound archive), and seven of the country's eight State and Territory libraries. The contributions of partners vary from simply selecting material to be archived, to negotiating with publishers, to programming the harvester to initiate a capture. So far, all the gathered material is stored and managed by the NLA. It will be interesting to see how this responsibility develops; there is an argument for sharing responsibility for storing, preserving, and providing access more equally among our partners, but there is also an argument that it is more efficient and reliable to centralize the storage and preservation functions.

In place of the inefficient harvesting and storage tools originally used, the Library has developed its own Digital Archiving System. This suite of tools has increased the efficiency of operations and made it easier for our partners to participate via a Web interface.

From the beginning, the Library has taken a selective approach to archiving. We believe the reasons for having taken this approach still apply. First, by archiving selectively we are able to focus some resources on quality control. We check each title to ensure that it has been copied completely and with full functionality (as far as is currently possible). Because all publications have been selected for their national significance and long-term research value, we consider this investment of time to be justified.

Second, by archiving selectively we can negotiate with publishers for permission to archive their publications (necessary in the absence of legal deposit legislation for digital publications) and make them accessible online or through dedicated onsite PCs.

While the Library recognizes many advantages in taking a more comprehensive approach to Web archiving, we have yet to be convinced that such advantages outweigh the benefits of quality control and accessibility that we have been able to achieve only while collecting selectively. However, we do not see these approaches as mutually exclusive. We would like to be able to pursue high-quality, ongoing capture for a core body of material selected for its research value, complemented by periodic capture of more comprehensive snapshots of the Australian domain.

Last year we engaged a consultant to look at the feasibility of such an approach. While funding difficulties interrupted this work, we are exploring a number of ways of making our national collection both broad and deep.

Although support for the enactment of legal deposit legislation for electronic publications is emerging, we will still need to communicate with many publishers to negotiate periods of restricted access

or assistance with formats that are difficult to gather automatically. For this reason, the NLA is working with Australian publishers to establish a code of practice that would guide us, particularly in dealing with commercial online publications.

PANDORA encounters a number of technical problems, even in its collection-building tasks. An early decision that content should take precedence over format means that, in principle, no publication is excluded simply because it is difficult to capture or manage. This is a noble objective but one that has not always been successful in practice. Despite many years' experience in automating our archiving processes, we still have to handcraft some features, such as applets, to make them work reliably. Our greatest difficulty is with publications structured as databases, which we have been unable to harvest. We plan to do more work in this area because we recognize the difficulty with databases as a major deficiency.

The ultimate purpose of all this effort is improved access. The NLA has long been interested in persistent identification that will keep information resources findable while they are still available. We are currently using an in-house system of persistent identifiers and resolution mechanisms.

Rights management is critical to PANDORA's access arrangements. While the archive has been developed to respect and support rights management for all publications, special procedures and controls have been developed for commercial publications.

From this discussion of collection building for PANDORA, it should be evident that our archiving arrangements continue to evolve as we encounter and deal with a wider range of issues.

Digital Preservation Programs

The Library's digital preservation programs have developed more slowly than has our digital collection building. However, some concrete steps are starting to emerge. Our preservation programs are predicated on a concern to protect and maintain the data stream carrying the archived information, and to maintain, and if necessary to recover, a means of accessing the archived information.

PANDORA

Our six-year experience in active archiving has taught us that the browsers that provide access to online material are remarkably tolerant. It is hard to find any material that cannot be accessed once it has been saved to the archive. This will change, especially as dependencies such as plug-in software are superseded and lost from users' PCs.

The most notable step we have taken with PANDORA has been to design and carry out a trial migration of files affected by the superseding of formatting tags in the HTML standard. Our modest migration does not constitute absolute proof that we can preserve access to the entire archive this way. However, it does suggest that we can quite efficiently make consistent, well-documented changes

within files in the archive and produce an outcome that meets our standards for preserving the significant properties of HTML files.

Physical Format Digital Publications

Our collection of physical format digital publications is not large; it comprises only a few thousand titles. It does, however, contain important material. In working with this collection, our most significant step has been to establish an ongoing regime of transferring information from unstable diskettes to more stable CD-Rs. We are about to experiment with transfer to a mass storage system. These are, again, quite minimal preservation steps aimed at enabling future preservation action.

Audio

The Library's collection of more than 35,000 hours of recorded sound has been a nursery for developing our thinking about digital preservation. We began digital recording in the early 1990s but did not begin archiving to a digital format until 1996. Our first archival digital carrier was CD-R, chosen for its manageability and expected reliability over the reasonably short time we intended to retain it. While always expecting to lose access to professional analog audio technology, until recently we have archived to both CD-R and analog tape. A few months ago, we finally dropped the analog part of our archiving strategy and moved to a digital mass storage system, managed through extensive metadata. This has been a rapid development over only five to six years, and most of the collection remains on analog tape in a controlled-climate store. We expect to be using our third or fourth mass storage system before we have copied the entire collection to a digital format.

Because the Library has retained control over the file formats we use and the quality of sound archiving work, we expect to use a straightforward migration path to maintain access to this collection.

Data Recovery

There is insufficient space in this paper for a detailed description of our data recovery program for the many undocumented diskettes that emerge from the Library's manuscript collections. Our investments in buying format recognition and translation software, and in developing procedures for using it, have been rewarded by regaining access to some important material (and to quite a lot of junk).

While not prepared to rely on data recovery as a means of ensuring ongoing access, we have come to accept it as a satisfactory method of last resort.

Infrastructure

Within Australia, it is likely that NLA's efforts in building infrastructure to manage all of these digital collections will be as seen as more important than the original initiatives themselves.

The following six types of infrastructure have been important for the Library:

1. tools
2. policy frameworks
3. resources (including expertise)
4. mechanisms for sharing information
5. collaborative agreements
6. certification

Tools

The key systems infrastructure to support all of our digital collections, and to carry and support the National Collection of Online Publications, is what we call our Digital Services Project. A challenge for the Library has been the lack of systems that could be bought off the shelf. Through procurement exercises starting in 1998, we tried to purchase systems for digital archiving, storage, and digital object management. Of the three, we have managed to buy only the storage system; the other two have had to be developed in-house—a slow and resource-intensive process.

Development of our systems was slowly aligned with the Open Archival Information System (OAIS) Reference Model, which is emerging as a standard framework (CCSDS 2001). The Library began modeling its business processes and data structures before we were aware of OAIS, and we continued to do so without feeling the need to fully adopt OAIS terminology. This apparent willfulness on our part does not seem to have caused either the NLA or anyone else much harm. At the right time, we found we could map our processes quite accurately to OAIS, providing something like an independent endorsement of the OAIS Reference Model.

Another essential tool for managing our digital collections is preservation metadata. Because we could find no existing model that met our needs, in 1999 we undertook development of a preservation metadata model to support all our digital collections. That work has contributed to the efforts of Research Libraries Group (RLG) and OCLC to negotiate a consensus metadata model that could be offered to the world (OCLC 2002).

Policy Frameworks

Policy is the second kind of infrastructure we needed to establish. In developing its Digital Services Project, the Library produced various information papers that serve as policy documents for many of our collection management processes.

More recently, we have set down our intentions regarding ongoing maintenance by releasing a digital preservation policy that addresses the way we manage our own collections as well as the way we wish to work with others (NLA 2002). Hindsight will probably see it as an early and rather unsatisfactory draft, but for now it is having a powerful effect in focusing our preservation efforts.

Like all good policies, this one has spawned an action plan that commits the Library to the following steps:

- Documenting our collections so that we understand what we have and what we have to deal with;
- Understanding and auditing the preservation effects of the ways we manage our collections currently;
- Developing mechanisms to monitor threats, preferably in collaboration with others;
- Defining the significant properties of our collections that must be maintained through our preservation processes; and
- Investigating how we can retain access to software required by our collections, and continuing our practical tests of emulation, migration, and other strategies, on the assumption that we will need to apply different approaches to different kinds of material. For example, we are confident that migration will work for our large, homogeneous collections of digital audio and image surrogates from our digitization programs, whereas emulation will probably be needed for parts of our physical formats collection and PANDORA, supported by ongoing access to software archives. We are also looking at the practicalities of using XML as a format simplification approach and at the use of generic document viewers for nonexecutable files, as currently used for our corporate electronic records.

Resources

While the NLA has been pleased to discover what it could achieve without outside funding, it would be foolish to deny that digital archiving and preservation programs are resource-intensive. The Library has had to reallocate quite a few million dollars from other work to achieve what it has been able to do so far in digital preservation. This reallocation has not been without pain, as the Library continues to acquire nondigital collections as rapidly as ever and remains as committed as it ever has been to their good management and preservation. The Library is reaching a point where it will be difficult to make further progress with its digital collections without additional resources and a sustainable business model.

With regard to managing workflows, the Library's practice of placing dedicated teams of specialists inside existing organizational units has proved effective in building expertise we require without losing contact with the broader institutional culture and direction.

Mechanisms for Sharing Information

We see information sharing as a critical enabler of digital preservation. The most visible manifestation of the Library's commitment to information sharing is PADI, the Web-based subject gateway on preserving access to digital information.

PADI was set up by a group of institutions as a place where we could share, compare, and find information about digital preservation. PADI is not the only good place to go looking, but our friends tell us it is their international subject gateway of choice. While managed by NLA, PADI has a number of contributors and partners, and a recent agreement with the Digital Preservation Coalition in the

United Kingdom ensures we will be working together to the benefit of both organizations' users.

Support from the Council on Library and Information Resources (CLIR), has allowed the NLA to pursue an experimental program of identifying and protecting the key resources listed in PADI through the Safekeeping Project. This project is based on a model of extremely distributed management of information resources, principally through self-archiving in compliance with a set of guidelines.

Collaborative Agreements

The NLA is committed to working with others in libraries, archives, universities, publishers, government agencies, and elsewhere, both in Australia and overseas. We seek to work collaboratively because our own small steps will take us only part of the way we need to go. We have found that collaboration works best where there is concrete action to be taken and clearly defined expectations on all sides.

Certification

It has long been recognized that some kind of certification is required to establish whether archiving arrangements can be trusted to provide adequate preservation guarantees (Task Force 1996; RLG 2002). The long history of cooperation between libraries in Australia may well lead us to look for cooperative ways of demonstrating our mutual accountability. It will be fascinating to watch the development of approaches to certification in other countries with different traditions of cooperation.

A National Model

In thinking about how national models for digital archiving may develop, it is helpful to return to the principles of responsibility mentioned earlier and the impact they have had in Australia.

- "Everyone doesn't have to do everything."
  This principle has made it possible for partners to come into PANDORA at a modest level of involvement. It has also allowed some people who do not have an identifiable role to opt out of active archiving.
- "We don't have to do everything at once."
  This message has enabled us to focus on collection building for the moment and to look for ways of improving how we manage collections later. It has also helped us accept the constraints and compromises along the way without falling into despair.
- "Responsibility can be time constrained."
  This principle has been especially powerful in inviting people to play a role for a defined period without implying a long-term obligation. It also helps us bear in mind that all of our roles may be time constrained and that effective exit strategies and succession plans are essential.

These principles have been useful in helping us approach and develop the building of a national model for distributed digital archiving. However, we believe that they are only valid in the context of some other related principles:

- "We may not all have to do everything, but someone has to do something."
- "Someone must be willing to take a lead on almost all steps."
- "In the last resort, someone must be willing to take responsibility for everything, even if it is only responsibility for a final decision that some information will be lost."

So far, building this national collection has worked well in Australia's library sector. That may have something to do with the NLA's leadership and the strong spirit of cooperation engendered by success. Perhaps out of success in individual sectors, it will be possible to achieve success within other sectors and among sectors, so that we can build a truly national model for archiving and preserving digital information.

References

All URLs were valid as of July 10, 2002.

Cathro, Warwick, Colin Webb, and Julie Whiting. 2001. Archiving the Web: The PANDORA Archive at the National Library of Australia. Available at http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html.

Consultative Committee for Space Data Systems. July 2001. Draft Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS). Available at http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf.

Fullerton, Jan. 1998. Developing National Collections of Electronic Publications: Issues to Be Considered and Recommendations for Future Collaborative Actions. Available at http://www.nla.gov.au/nla/staffpaper/int_issu.html.

Library of the University of California, Berkeley. 1996. Digital Library SunSITE Collection and Preservation Policy. Available at http://sunsite.berkeley.edu/Admin/collection.html.

National Library of Australia. 1997. Statement of Principles for the Preservation of and Long-Term Access to Australian Digital Objects. Available at http://www.nla.gov.au/preserve/digital/princ.html.

National Library of Australia. 2002. A Digital Preservation Policy for the National Library of Australia. Available at http://www.nla.gov.au/policy/digpres.html.

OCLC Online Computer Library Center. 2002. OCLC/RLG Preservation Metadata Working Group. Available at: http://oclc.org/research/pmwg/.

Research Libraries Group and Online Computer Library Center. 2002. Trusted Digital Repositories: Attributes and Responsibilities. Available at http://www.rlg.org/longterm/repositories.pdf.

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* Washington, D.C.: Commission on Preservation and Access, and Mountain View, Calif.: Research Libraries Group. Available at http://www.rlg.org/ArchTF/.

Web sites noted in paper:

National Library of Australia Digital Services Project: http://nla.gov.au/dsp

PADI: http://www.nla.gov.au/padi

PANDORA: http://pandora.nla.gov.au/

Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information

Donald Waters

"Good fences make good neighbors." This famous aphorism from Robert Frost's poem "Mending Wall" suggests the title and subject of this paper.[1] Let me begin by explaining the relevance of the poem to the topic of the archiving of digital information.

A Preservation Parable

"Mending Wall" is a parable in the form of a poem. Wonderfully crafted, it can be read on many levels. It is about boundaries and territoriality, the conflict between primitive impulse and modern reflection, and the nature of ritual and work.[2] But at another level, "Mending Wall" is simply about the preservation of a shared resource—a common wall that each year two neighbors must join together to rebuild. Why does it need repair? As the opening line famously puts it, "Something there is that doesn't love a wall."

The narrator of the poem identifies two sources of damage: natural causes, such as the heaving of stones that results from the freezing and thawing of the earth, and deliberate human acts, such as the attempts of hunters and their dogs to flush out their prey from hiding in the wall. Whatever the cause, it is the mending that matters to the narrator, who says:

> The gaps I mean,
> No one has seen them made or heard them made,
> But at spring mending-time we find them there.
> I let my neighbor know beyond the hill;
> And on a day we meet to walk the line
> And set the wall between us once again.

[1] For the poem, see Lathem (1979: 34-35).

[2] For critical commentaries on the poem, see Nelson (2001) and Faggen (2001).

The narrator then vividly describes the mending process. The fieldstones are heavy and variously shaped; they often do not fit well together. He says, "We wear our fingers rough with handling them." All of this is a hard, but straightforward, technical process. Then the neighbors come to a grove of trees, and the narrator asks why do we need to mend the wall here? The taciturn New England reply of the neighbor is simple: "Good fences make good neighbors."

From this point, the poem takes a darker turn as the conflict between the narrator and his neighbor becomes apparent. The narrator probes deeper into the reasons why neighbors agree to preserve their common resources. "Before I built a wall I'd ask to know/What I was walling in or walling out,/And to whom I was like to give offense." But the neighbor's motives remain inscrutable.

> I see him there,
> Bringing a stone grasped firmly by the top
> In each hand, like an old-stone savage armed.
> He moves in darkness as it seems to me
> Not of woods only and the shade of trees.

The neighbor simply will not admit that letting the wall deteriorate is a possibility and says again to conclude the poem, "Good fences make good neighbors."

And so the reader is left with a puzzle. The wall has different meanings to each of the neighbors and, although the narrator calls his neighbor each year to the task, he himself finds many reasons to question the merits of preserving the wall. So what moves these two people to come together each year to mend this common resource? Could it be that what makes good neighbors is not simply a boundary? Could it be that what makes good neighbors is the very act of keeping the common resource good—of making and taking the time together to preserve and mend it?

The Archiving of Digital Information

The library, publisher, and scholarly communities are now engaged in efforts to resolve the problems associated with preserving another kind of common resource: digital information. Such information is a critical priority, especially for libraries and other institutions that have borne responsibility for maintaining the cultural record. Six years have now passed since the Task Force on Archiving of Digital Information issued its report (Waters and Garrett 1996). During the course of its work from 1994-1996, the Task Force recognized well that "something there is that doesn't love digital information." In the face of the limits of digital technology, the Task Force struggled, as does Frost's narrator, with the question of motivation and action: Why should we preserve digital information, and who should do it?[3]

[3] The word "archiving" has multiple senses ranging from the narrow sense used by professional archivists to designate the process of preserving formal records to the broad sense used by computer technologists to refer to a temporary backup collection of computer files. For certain purposes and audiences, one might choose to restrict use of the word to one or other of these senses. In this paper, I have followed the Task Force on Archiving of Digital Information (Waters and Garrett 1996), and use digital archiving and digital preservation interchangeably to refer to the long-term maintenance of digital objects judged to be of enduring value.

The Task Force's response was that we need a serious investment in archiving because we are in danger of losing our cultural memory. The first line of defense rests with creators, providers, and owners, who must take responsibility for creating archivable content. A deep infrastructure, consisting of trusted organizations capable of storing, migrating, and providing access to digital collections, is then needed for the long term. A process of certification to establish a climate of trust is also needed, as is a fail-safe mechanism by which certified archives would have a right and duty to exercise aggressive rescue of endangered or "orphaned" materials.

Since the Task Force report was issued, there has been much experimentation, definition of requirements, and development, much of it reported and summarized in previous papers in this volume.[4] Margaret Hedstrom has reported the recent emergence of a greatly sharpened sense of the research needed to support digital archives. The development of the Reference Model for Open Archival Information Systems (OAIS) has been a galvanizing force (CCSDS 2001). As Titia van der Werf, Colin Webb, and others have described, a number of digital archives have been created, are being created, or are expanding following the OAIS model in the United States, the United Kingdom, the European Union, and Australia.[5] Most of these efforts, however, have been government-funded, a point to which I return below.

In other developments, the emulation-versus-migration debate has largely played itself out. Neither approach provides a sufficient, general answer to the problem of digital preservation, and it has proven largely fruitless to debate the merits of these approaches in the abstract.[6] Instead, there is growing recognition that different kinds of information captured in different ways for long-term preservation will need various kinds of support.

Thanks to a variety of reports, such as those organized by the Research Libraries Group (RLG) and the Online Library Computer Center (OCLC) on preservation metadata for digital objects and attributes of trusted digital repositories, there is a deepening understanding of the requirements and expectations for best practices when building trustworthy archives.[7] Some of these analyses of requirements, it must be noted, are also being conducted in the abstract, without a realistic sense of costs and what will work, and so may be setting unrealistic expectations. Nevertheless, much is being learned from all these initiatives.

---

[4] The Preserving Access to Digital Information (PADI) Web site, which is maintained by the National Library of Australia, is one of the most comprehensive and up-to-date sources of information about the archiving of digital information. Available at: http://www.nla.gov.au/padi/.

[5] For a recent overview, see also Hodge and Carroll (1999).

[6] See, for example, the largely polemical debate on the relative merits of emulation and migration in Rothenberg (1999) and Bearman (1999). For a more balanced view, see Granger (2000).

[7] See OCLC (2002) and RLG (2002). For a different approach to requirements definition, see Cooper, Crespo, and Garcia-Molina (2000).

Our vision is much less clear about the infrastructure needed to enable archives to cooperate and interoperate. Our understanding of the legal and business frameworks needed to sustain the long-term preservation of digital information is likewise still very crude.[8] For those interested in these questions, a recent initiative of the Mellon Foundation that was designed to explore the archiving of electronic journals may shed some light. This paper describes some of the results of that project, lays out some of the issues the participants have encountered, and suggests some solutions.

Mellon Electronic Journal Archiving Program

Over the last decade, there has been much hope placed in the potential of electronic publishing as a means of resolving the rising costs of scholarly publishing.[9] However, with the recent dot.com collapse has come an increasingly sober approach to electronic publishing. One aspect of the reassessment that is under way is a growing awareness that archiving has not yet been factored into the overall costs of the system, and if electronic publishing is to be taken seriously, it must be.

Given the general digital archiving problem, and the Foundation's particular concern with scholarly publishing, Foundation staff began several years ago consulting with librarians, publishers, and scholars about how best to stimulate investments in solutions. An investment in the archiving of electronic journals seemed to be especially promising and was welcomed by both publishers and libraries. The Foundation solicited proposals for one-year planning projects, and, in December 2000, the trustees selected seven for funding.[10] Building on what has been learned during these planning efforts, the Foundation is now preparing to fund two major implementation projects.

What was the reason for focusing on e-journals? Scholars demand the multiple advantages of this emerging medium, including reference linking, easy searching across issues and titles, and the ability to include data sets, simulation, multimedia, and interactive components in the published articles. In addition to flexibility and functionality, e-journals have promised lower costs, but this goal has remained elusive. Major journals are rarely published only in e-format, and the costs of archiving are unknown. Without trusted electronic archives, it is unlikely that e-journals can substitute for print and serve as the copy of record, and so we have a duplicative and even more costly system—a system we all hope is transitional.[11]

---

[8] For approaches to these topics, see, for example, Granger (2002), and Cooper and Garcia-Molina (2001).

[9] See, for example, Ekman and Quandt (1999).

[10] Copies of the successful proposals are available at http://www.diglib.org/preserve/ejp.htm. See also Flecker (2001). For another perspective on the archiving of electronic journals, see Arms (1999).

[11] For data on these issues, see, for example, Born and Van Orsdel (2001), and Van Orsdel and Born (2002).

Of the seven Foundation-funded planning projects, the Stanford University project proposed to develop a technology for harvesting presentation files—the Web-based materials that publishers use to present journal content to readers—and storing them in a highly distributed system called LOCKSS, (Lots Of Copies Keeps Stuff Safe). Five projects engaged in planning for the capture of publishers' source files, including high quality images and text that is encoded in the standard generalized markup language (SGML) or the extensible markup language (XML).[12] Three of these explored a publisher-based approach: Harvard worked with Wiley, Blackwell, and the University of Chicago Press; the University of Pennsylvania worked with the Oxford and Cambridge University presses; and Yale partnered with Elsevier. The two other projects took a discipline-based approach: Cornell focused on journals in agriculture, and the New York Public Library focused on e-journals in the performing arts. In the seventh project, the Massachusetts Institute of Technology explored the issues involved in archiving what it saw as a new class of periodical publication made possible by the digital medium—publications that it referred to as "dynamic e-journals." These publications included CogNET and Columbia International Affairs Online (CIAO).[13]

When inviting proposals for these projects, the Foundation asked applicants to focus on a rather complicated set of objectives. They were asked to:

- identify publishers with which to work and to begin to develop specific agreements regarding archival rights and responsibilities
- specify the technical architecture for the archive, perhaps using a prototype system
- formulate an acquisitions and growth plan
- articulate access policies
- develop methodologies to be used to validate and certify the repository as a trusted archive
- design an organizational model, including staffing requirements and the long-term funding options, that could be tested and evaluated during a setup phase

These were ambitious goals, and the outcomes that the Foundation trustees expected were equally ambitious. They hoped that leading research institutions, in partnership with specific publishers,

---

[12] In what follows, I distinguish two approaches to archiving: one that focuses on the capture of presentation files; the other that focuses on source file capture. Dale Flecker of Harvard University points out, in a personal communication dated May 30, 2002, that for many publishers, SGML or XML files are not really source files, but are among a variety of derivative files that are generated during the publication process. Referring to a "source file approach" to electronic journal archiving thus may be inaccurate from at least one perspective. I have nevertheless retained the label because the intent of this group of planning projects was to identify and capture files that would serve both publishers and archives as an authoritative source from which e-journal content could be reliably disseminated to a reader as the technology for representation and display changed over time.

[13] See http://www.cognet.org/ and http://www.ciaonet.org/.

would develop and share detailed understandings of the requirements for setting up and implementing trustworthy archives for electronic journals; that enabling technology would be developed to facilitate the archiving process; and that plans would be developed as competitive proposals designed to secure funding for the implementation and operation of electronic journal archives.

The planning period has come to an end, and much has been accomplished. In this paper, I cannot analyze how each of the projects succeeded or failed in meeting the ambitious goals and expectations set for them.[14] Instead, I would summarize the findings by noting, first, that archiving now seems technically feasible using different approaches: the capture of Web-based presentation files using LOCKSS and the capture of source files. Second, participating publishers have come to view archiving their journals as a competitive advantage. Third, there is an increasingly shared understanding that an e-journal archive should aim to make it possible to regard e-journals as publications of record and to persuade publishers and libraries to consider abandoning print. There were other key results, some of them unexpected. I now turn to a discussion of the most important of these, which relate to the economics and organization of digital preservation.

The Political Economy of Public Goods

In trying to devise next steps, the project teams ran smack into some of the classic problems of the political economy of public goods—questions that Robert Frost explored in a much more elegant and artful way. What are the incentives for individuals and institutions to participate in the provision of a good from which others cannot be readily excluded from enjoying the benefit? What are the organizational options? What are sustainable funding plans?

The Task Force on Archiving of Digital Information argued that the value of digital information rests in what it contributes to our cultural memory. Because cultural memory is a public good, it follows that insuring against the possible loss of such memory by the archiving of digital information would also be a public good. The joint economic interest of publishers, authors, and the scholarly community in electronic journals as intellectual property is reason to suggest that archiving them may not be a public good in the strictest sense of the term. Still, the archiving of digital information has special properties as a kind of modified public good that demands special attention.[15]

---

[14] Each of the institutions that participated in the Mellon Electronic Journal Archiving Initiative is preparing a final report of its planning project. All reports should be available by September 2002 at http://www.diglib.org/preserve/ejp.htm.

[15] For a strict definition of a public good, see Baden (1998: 52): "A public good is one which, if available for anyone, is available for everyone. . . . This suggests that the good is not easily packaged for sale, and people cannot be excluded from its consumption. In other words, property rights cannot be readily established for public goods. A public good is also one whose incremental use does not reduce, subtract, or consume it."

To understand these properties, let us examine the proposition that archiving is insurance against the loss of information. Is archiving really like insurance, in the sense of life or fire insurance? Would a business model for archiving based on an insurance model induce people to take on responsibility for archiving? If you have fire insurance and your house burns down, you are protected. If you have life insurance and you die, your heirs benefit. There is an economy in these kinds of insurance that induces you to buy. If you fail to buy, you are simply out of luck; you are excluded from the benefits. Unfortunately, the insurance model for archiving is imperfect, because insurance against the loss of information does not enforce the exclusion principle.[16]

A special property of archiving is that if one invests in preserving a body of information and that information is eventually lost to others who did not take out the insurance policy, the others are not excluded from the benefits, because the information still survives. Because free riding is so easy, there is little economic incentive to take on the problem of digital preservation, and this partly explains why there has been so little archive building other than that funded by governments. Potential investors conclude that "it would be better for me if someone else paid to solve the archiving problem." In fact, one of the defining features of a public good—and think here of other public goods such as parks or a national defense system—is that it is difficult and costly to exclude beneficiaries.

The Tragedy of the Commons

Given the huge free-riding problem associated with the maintenance of public goods, what are the alternatives? Reflecting in part on this problem, Garrett Hardin in an influential article entitled "The Tragedy of the Commons," despaired of solutions. "Ruin," he wrote, "is the destination toward which all men rush, each pursuing his own interest in a society that believes in the freedom of the commons. Freedom in a commons brings ruin to all" (1968, 1244). Hardin echoed Thomas Hobbes, who lamented the state of nature, a commons in which people pursue their own self-interest and lead lives that are "solitary, poore, nasty, brutish, and short" ([1651] 1934, 65). Remember the state-of-nature allusion in the Frost parable about preserving a common resource? To the narrator, the neighbor seems "like an old-stone savage armed."

Focused on preserving digital information in 1996, the Task Force on Digital Archiving echoed both Hobbes and Hardin in writ-

---

[16] There is a substantial literature on the economics of various types of insurance, which is broadly defined as a mechanism that "mitigates against the influence of uncertainty" (McCall 1987: 868). For analyses of the problems in creating markets for insurance, see, for example, Arrow (1963), Pauly (1968), Ehrlich and Becker (1972), and Hirshleifer and Riley (1979).

There may be great utility in viewing digital preservation in terms of the business of insurance with its apparatus of risk management and underwriting. Some preliminary and promising applications of the economics of insurance to the problems of digital archiving include Lawrence (1999) and Kenney (2002).

ing that "rapid changes in the means of recording information, in formats for storage, in operating systems, and in application technologies threaten to make the life of information in the digital age 'nasty, brutish, and short'"(Waters and Garrett 1996, 2). One of Hardin's solutions to the tragedy of the commons was, like Hobbes's, to rely on the leviathan—the coercive power of the government. Certainly, protection of the common good in the archiving of digital information could be achieved by massive government support, perhaps in combination with philanthropy.

Given these considerations of public goods economics, it is no accident that so many of the existing archiving projects are government funded, and it may be that some forms of archiving can be achieved only through a business model that is wholly dependent on government or philanthropic support. Several national governments, including our own through the agency of the Library of Congress, are exploring the power of copyright deposit and other mechanisms for developing digital archives. The National Archives and Records Administration is financing major archiving research projects with the San Diego Supercomputer Center and other organizations. Brewster Kahle's Internet Archive, which has been collecting and storing periodic snapshots of the publicly accessible Web, is an extraordinary example of philanthropic investment in digital archiving by someone who made his fortune in the development of supercomputers.[17]

Hardin's other solution to the tragedy of the commons was to encourage its privatization, trusting in the power of the market to optimize behavior and preserve the public good. It is not unreasonable to view congressional extensions of copyright and other measures to protect the rights of owners as efforts to privatize intellectual property and entrust its preservation to the self-interest of owners.[18] Advocates of author self-archiving articulate a similar trust of self-interest in the service of the public good.[19] Moreover, in the digital realm, as with other forms of information, the passions and interests of what Edward Tenner has called "freelance selectors and preservers" will almost surely result in valuable collections of record (2002, 66). Just as government and philanthropy undoubtedly have a role in digital archiving, so too will private self-interest. In fact, the Task Force report suggested that the first (but not last) line of defense in digital archiving rests with creators, providers, and owners.

Organizational Options

Government control and private interest, however, are unlikely to be sufficient, or even appropriate in many cases, for preserving the public good in digital archiving. Moreover, substantial experimental and

[17] See http://www.archive.org/.

[18] Whether such extensions are good public policy is the subject of vigorous debate. See, for example, Lessig (2001) and Vaidhyanathan (2001).

[19] See, for example, Harnad (2001).

field research in the political economy of public goods has shown Hardin's pessimism about the prospects of maintaining public goods to be unwarranted. Case after case compiled since Hardin published in 1968 demonstrates that groups of people with a common interest in a shared resource will draw on trust, reciprocity, and reputation to devise and agree upon rules for and the means of financing the preservation of the resource.[20] The projects that Mellon funded provide seven more case studies with similar prospects for e-journal archiving.

The Mellon Foundation will undoubtedly continue to pursue its long-standing philanthropic interest in the preservation of the cultural record as a condition of excellence in higher education. At the same time, it is looking, as it does in nearly all cases of support, for ways to promote a self-sustaining, businesslike activity. It seeks to foster the development of communities of mutual interest around archiving, help legitimize archiving solutions reached within these communities, and otherwise stimulate and facilitate community-based archiving solutions. The premise of the Mellon e-journal projects was that concern about the lack of solutions can be addressed only by hard-nosed discussions among stakeholders about what kinds of division of labor and rights allocations are practical, economical, and trustworthy.

What about publisher-based archives? The question here is not whether preservation is in the mission of publishers. As long as their databases are commercially viable, publishers have a strong interest in preserving the content—either themselves or through a third party. Scholarly publishers also have an incentive to contribute in the interests of their authors, who want their works to endure, be cited, and serve as building blocks for knowledge. However, the concern about the viability of publisher-based archives is whether the material is in a preservable format and can endure outside the cocoon of the publisher's proprietary system. One necessary ingredient in a proof of archivability is the transfer of data out of their native home into an external archive, and as long as publishers refuse to make such transfers, this proof cannot be made.

The research libraries of major universities are also interested, some say by definition, in ensuring that published materials are maintained over the long term. With regard to the digital archiving of electronic journals, the libraries in the Mellon projects have generated several significant technical and organizational breakthroughs. They demonstrated that digital archiving solutions that meet the needs of the scholarly community require at least three factors: extreme sensitivity to public goods economics, dramatic efforts to take advantage of the economies of scale inherent in the technology either through centralization or a radical distribution of service, and very low coordination costs in consistently and transparently managing publisher and user relations. Meeting these requirements within existing library structures has proved elusive, but in mapping out what

[20] See, for example, Ostrom (1990), Bromley (1993), Anderson and Simmons (1993), Baden and Noonan (1998), and Ostrom (1999).

these requirements are, some of the most imaginative minds working in libraries today have blazed trails in the Mellon-sponsored projects and demonstrated what solutions are likely to succeed in the next phase of the e-journal initiative.

What Would Be the Economic Model?

One of the surprising findings that the Mellon Foundation has made in monitoring these projects is that new organizations are likely going to be necessary to act in the broad interest of the scholarly community and to mediate the interests of libraries and publishers. But if some new archival organization (or organizations) were created to perform the preservation function, what rights and privileges would they need to be able to sustain the e-journal content? Can ways be found to apply the exclusion principle in such a manner that it creates an economy for digital archiving—a scarcity that publishers and libraries are willing to pay to overcome and that would support the larger public good? Put another way, what kinds of exclusive benefits can be defined to induce parties to act in the public good and invest in digital archiving?

Access is the key. Over and over again, we have found that one special privilege that would likely induce investment in digital archiving would be for the archive to bundle specific and limited forms of access with its larger and primary responsibility for preservation. User access in some form is needed in any case for an archive to certify that its content is viable. But extended and complicated forms of access not only add to the costs of archiving, they also make publishers very nervous that the archives will in effect compete for their core business. As a result, the Foundation is now looking to support models of archival access that serve the public good but that do not threaten the publishers' business.

Secondary, noncompeting uses might include aggregating a broad range of journals in the archive—a number of publications larger than any single publisher could amass—for data mining and reflecting the search results to individual publishers' sites. Another kind of limited, secondary use might be based on direct access to the content with "moving walls" of the kind pioneered in JSTOR.[21] Much work needs to be done to sort out what the right access model might be, but it is clear that so-called "dark" archives, in which a publisher can claim the benefit of preservation but yields no rights of access, do not serve the public good. They serve only the publisher, and the Foundation is not willing to support such archives.

Archiving requires agreements. The basic value proposition for digital archiving that has thus emerged from these projects is this: Publishers would bear the costs of transferring their content in an archivable form to a trusted archive and allow a limited but significant form of access or secondary use as part of the archiving process. Uni-

[21] See http://www.jstor.org/about/movingwall.html.

versities and colleges, through their libraries, would pay for the costs of preservation in exchange for a specific but limited form of access; those who do not contribute do not get the access. Given this form of participation by publishers and universities, e-journal archives would maintain the content over time. This bargain would have to be cemented organizationally and legally in the form of appropriate licenses that define in detail what content is archived, the responsibilities of the parties, and the conditions of use.

Priming the Pump

To prime the pump for such self-sustaining, community-based solutions for the archiving of scholarly electronic journals, the Foundation is now focused on developing support for the two approaches explored in the planning process just concluded, namely, preserving presentation files using LOCKSS and preserving source files.

Preserving presentation files with LOCKSS. In the LOCKSS system, a low-cost Web crawler is used for systematically capturing presentation files. Publishers allow the files to be copied and stored in Web caches that are widely distributed but highly protected. The caches communicate with each other through a secure protocol, checking each other to see whether files are damaged or lost and repairing any damage that occurs. Caching institutions have the right to display requested files to those who are licensed to access them if the publisher's site is unavailable and to provide the local licensed community the ability to search the aggregated files collected in the institutional cache.

During the next phase of development, the key issues for the LOCKSS system are to separate the underlying technology from its application as an e-journal archiving tool; explore ways of ensuring the completeness and quality of e-journal content on acquisition and of managing the content as bibliographic entities rather than simply as Web-addressed files; expand the coverage of journals; maintain the LOCKSS software; and identify strategies for migrating the e-journal content. To help undertake and finance these tasks, Stanford has identified a variety of partners and is planning the development of a LOCKSS consortium.

Preserving source files. The source file capture approach requires that publishers be able to present, or "push," files in a normalized form to the e-journal archive. The question of cost in this approach turns, at least initially, on how many output formats a publisher must produce and how many an archive must support from different publishers. During the course of its project, Harvard commissioned a consultant's report to determine the feasibility of developing a standard archival interchange document type definition (DTD) that would dramatically reduce this complexity (Inera 2001). The report suggests that it is possible to produce such a DTD without reducing content to the lowest common denominator, sacrificing substantial functionality and appearance, or avoiding attention to extended character sets, mathematical symbols, tables, and other features of

online documents. The planning projects also made significant progress in specifying both the tools needed to transfer, or "ingest," e-journal content into the archive and a workflow for managing content quality control. License agreements were also outlined that began to converge on the concepts of "moving walls" and other limited rights of user access.

What are the next steps for developing the source file capture approach? The cost and scale of archiving source files suggest the need for a coordinated and collaborative approach for shaping the agreements with publishers, developing the underlying archival repository, and creating operational procedures for transferring content from publisher to archive. One approach that the Foundation is considering would be to channel the expertise and energy developed in these projects through a not-for-profit entity that is either part of JSTOR or related to it. Such an entity would be expected to assume archiving responsibility for a substantial subset of the electronic journal literature for the academic community and would require investment by the university community to obtain the benefits of secondary *access* rights that the archive would provide and that would not compete with the core business of the publishers. This is not to say that the business model and terms of participation that currently exist at JSTOR are a perfect fit for electronic archiving, but rather that a lean, entrepreneurial, mission-driven organization such as JSTOR, which is positioned at the nexus of publishers, libraries, and scholars, is well situated to take the development of the archive to the next step.[22] As the new organization begins to take shape, the Foundation expects to involve the participants from the planning projects, to incorporate the specific breakthroughs each participant has made, and to think about the specific models of access and cost recovery that would be necessary to preserve and sustain electronic journal content for the common good of the scholarly community.

These two approaches are very different. Although experience might later tell us that one approach is better suited than the other for certain kinds of materials, it would not be useful now to think of them as competing approaches. We have to get used to the idea that overlapping and redundant archiving solutions under the control of different organizations with different interests and motives in collecting offer the best hope for preserving digital materials. We currently have no operating archives for electronic journals. It would be unwise at the outset to expect that only one approach would be sufficient.

Moreover, these different approaches suggest a natural layering of functions and interfaces from the repository layer to access services. Given such points of interaction, specialization and division of labor are possible that could result in real economies. If there are economies of scale in the LOCKSS system, for example, some functions could be more centralized in what was conceived as a highly

---

[22] JSTOR has developed significant expertise in the archiving of electronic journals that could be greatly leveraged. See Guthrie (2001). For a further account of JSTOR's archiving activities, see the presentation by Eileen Fenton at http://www.jstor.org/about/e.archive.ppt.

decentralized system. Conversely, source file capture could make greater use of distributed storage. Possibilities exist for even further development. Files aggregated in the archives across publishers could serve secondary abstract and indexing publishers as a single source, not only saving them from going to each and every publisher for the texts to index but also enabling them to use computational linguistic and other modern techniques to improve their products. Source files might also be "born archival" at the publisher and deposited in the archive, from which they might then serve as the masters for the derivative published files that the publisher creates for its different markets. These latter possibilities are not likely to emerge immediately, mainly because they would require intense negotiation among the interested parties; however, they are suggestive of how a thoughtful, entrepreneurial, community-based approach to archiving might add incremental improvements that would actually lead to more dramatic transformations of the system of scholarly communications.

Broader Context and Conclusions

The approaches to e-journal archiving that the Foundation and its partners are now considering would have to be formulated in the context of a much broader array of solutions for the archiving of digital information. An especially important part of this larger context is the development of local institutional archives for the variety of scholarly digital materials that members of each college or university community create but have little means of maintaining over time. The basis for what appears in scholarly journals will undoubtedly be found in data sets and other supporting materials at the authors' home institutions. In addition, a range of archival solutions needs to be developed for the much broader array of digital content in the form of newspapers, popular periodicals, music, video, scientific data sets, and other digital content that the cultural and scholarly community deems important for long-term preservation.

Another element in the larger context, and a critical impediment for digital archiving that arises again and again, is the legal regime governing intellectual property. There is now considerable confusion among policy makers in the United States about how the protections that have been afforded to owners of intellectual property in the digital age should serve to advance the higher goal established in the U.S. Constitution of promoting "the progress of science and useful arts."[23] For print materials, special exemptions have been built into the copyright law for preservation activities.[24] It may be too early to formulate specific exemptions that would apply to digital information. However, instead of waiting indefinitely for the policy confusion to be resolved, one step forward may be to begin to articulate

[23] U.S. Constitution, Article 1, Section 8, Clause 8.

[24] U.S. Code, Title 17, Section 108.

"safe harbor" principles about intellectual property rights that could form the basis of digital archiving agreements among interested parties. In building JSTOR and ArtSTOR, the Foundation has found that content owners are much more comfortable with agreements that limit uses of intellectual property to not-for-profit educational purposes than they are with agreements that leave open the possibility of creating competing commercial profit-making access to the property. Lawrence Lessig has also recently argued for the utility of the distinction between not-for-profit educational uses and other kinds of uses of intellectual property (2001, 249-261). Because educational use is certainly consistent with the Constitutional mandate for intellectual property law in the United States to promote "the progress of science and useful arts," perhaps it is time to build a safe-harbor framework for digital archiving on just such a distinction.

It is on this point that we come back to Robert Frost's preservation parable. I suggested earlier that what makes good neighbors may not be simply a boundary. Rather what makes good neighbors is the very act of keeping good the common resource between them—the act of making and taking the time together to preserve and mend the resource. So too it is with digital archiving.

In the context of an array of factors relating to many kinds of digital materials, the lessons of the Mellon planning projects are clear. Relevant stakeholders—scholars, publishers, and research libraries—can frame the archiving problem very concretely as a problem of technical, organizational, and economic development. Two options are being actively explored as a result. The first, LOCKSS, appears to be a relatively inexpensive solution, but caution is needed because the system may not be capturing files in the best long-term format. The second option, source file capture, is likely to be more expensive but promises to support the most durable archive. Framed in this way, using a variety of approaches, digital archiving, for electronic journals at least, seems achievable as what one might call a modified public good.

There are many dimensions to the good to be achieved, but two merit special mentioning. On the one hand, there is the joining together by scholars and the agents of education—universities, libraries, scholarly societies, and publishers—in serving the common interest of future scholarship by keeping good, or preserving, the digital resources now being created. On the other hand, there is the research and learning thereby made possible, which are the indelible marks of a good scholar. In other words, good archives make good scholars. If we accept the proposition that a free society depends on an educated citizenry, it is not a great leap of logic to conclude further that good archives make good citizens.

References

All URLs were valid as of July 10, 2002.

Anderson, Terry L., and Randy Simmons, eds. 1993. *The Political Economy of Customs and Culture: Informal Solutions to the Commons Problem.* Lanham, Md.: Rowman and Littlefield Publishers, Inc.

Arms, William. 1999. Preservation of Scientific Serials: Three Current Examples. *Journal of Electronic Publishing* 5 (December 1999). Available at: http://www.press.umich.edu/jep/05-02/arms.html.

Arrow, Kenneth J. 1963. Uncertainty and the Welfare Economics of Medical Care. *American Economic Review* 53 (December): 941-973.

Baden, John A. 1998. A New Primer for the Management of Common-Pool Resources and Public Goods. In Baden and Noonan (1998): 51-62.

Baden, John A., and Douglas S. Noonan, eds. 1998. *Managing the Commons*, 2nd ed. Bloomington: Indiana University Press.

Bearman, David. 1999. Reality and Chimeras in the Preservation of Electronic Records. *D-Lib Magazine* 5(4) (April). Available at: http://www.dlib.org/dlib/april99/bearman/04bearman.html.

Born, Kathleen and Lee Van Orsdel. 2001. Searching for Serials Utopia: Periodical Price Survey 2001. *Library Journal* (April 15): 53-58.

Bromley, Daniel, ed. 1993. *Making the Commons Work: Theory, Practice, and Policy.* San Francisco: ICS Press.

CCSDS, 2001. Reference Model for an Open Archival Information System (OAIS), Draft Recommendation for Space Data System Standards, CCSDS 650.0-R-2. Red Book. Issue 2. Washington, D.C.: Consultative Committee for Space Data Systems. July. Available at: http://www.ccsds.org/RP9905/RP9905.html.

Cooper, Brian, Arturo Crespo, and Hector Garcia-Molina. 2000. Implementing a Reliable Digital Object Archive. *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL).* Lisbon, Portugal, September 18-20. Available at: http://dbpubs.stanford.edu:8090/pub/2000-28.

Cooper, Brian, and Hector Garcia-Molina. 2001. Creating Trading Networks of Digital Archives. Delivered at the First ACM/IEEE Joint Conference on Digital Libraries. June 24-28. Roanoke, Virginia. Available at: http://dbpubs.stanford.edu:8090/pub/2001-23.

Ehrlich, Isacc, and Gary S. Becker.1972. Market Insurance, Self-Insurance, and Self-Protection. *Journal of Political Economy* 80 (July-August): 623-648.

Ekman, Richard, and Richard Quandt, eds. 1999. *Technology and Scholarly Communication.* Berkeley: University of California Press.

Faggen, Robert, ed. 2001. *The Cambridge Companion to Robert Frost.* New York: Cambridge University Press.

Flecker, Dale. 2001. Preserving Scholarly E-Journals. *D-Lib Magazine* 7(9) (September). Available at: http://www.dlib.org/dlib/september01/flecker/09flecker.html.

Granger, Stewart. 2000. Emulation as a Digital Preservation Strategy. *D-Lib Magazine* 6(10) (October). Available at: http://www.dlib.org/dlib/october00/granger/10granger.html.

Granger, Stewart. 2002. Digital Preservation and Deep Infrastructure. *D-Lib Magazine* 8(2) (February). Available at: http://www.dlib.org/dlib/february02/granger/02granger.html.

Guthrie, Kevin. 2001. Archiving in the Digital Age: There's a Will but is There a Way? *EDUCAUSE Review* (December): 56-65. Available at: http://www.educause.edu/ir/library/pdf/erm0164.pdf.

Hardin, Garrett. 1968. The Tragedy of the Commons. *Science,* 162 (December 13): 1244.

Harnad, Steven. 2001. The Self-archiving Initiative: Freeing the Refereed Research Literature Online. *Nature* 410 (April 26), 1024-1025. Available at: http://www.ecs.soton.ac.uk/~harnad/Tp/nature4.htm.

Hirshleifer, J., and John G. Riley. 1979. The Analytics of Uncertainty and Information—An Expository Survey. *Journal of Economic Literature* 17 (December): 1375-1421.

Hobbes, Thomas. [1651] 1934. *Leviathan.* Reprint. London: J. M. Dent & Sons, Ltd.

Hodge, Gail, and Bonnie C. Carroll. 1999. *Digital Electronic Archiving: The State of the Art and the State of the Practice. A Report Sponsored by the International Council for Scientific and Technical Information, Information Policy Committee, and CENDI.* Available at: http://www.dtic.mil/cendi/proj_dig_elec_arch.html.

Inera, Inc. 2001. *E-Journal Archive DTD Feasibility Study. Prepared for the Harvard University Library, Office of Information Systems, E-Journal Archiving Project.* Available at: http://www.diglib.org/preserve/hadtdfs.pdf.

Kenney, Anne, et al. 2002. Preservation Risk Management for Web Resources, *D-Lib Magazine* 8(1) (January). Available at: http://www.dlib.org/dlib/january02/kenney/01kenney.html.

Lathem, Edward Connery, ed. 1979. *The Poetry of Robert Frost: The Collected Poems, Complete and Unabridged.* New York: Henry Holt and Company.

Lawrence, Gregory, et al. 1999. *Risk Management of Digital Information: A File Format Investigation.* Washington, D.C.: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/pub93/contents.html.

Lessig, Lawrence. 2001. *The Future of Ideas: The Fate of the Commons in a Connected World.* New York: Random House

McCall, J. J. 1987. Insurance. In *The New Palgrave: A Dictionary of Economics.* Volume 2. Edited by John Eatwell, Murray Milgate, and Peter Newman. London: Macmillan. Pp. 868-870.

Nelson, Cary, ed. 2001. On Mending Wall. *Modern American Poetry: An Online Journal and Multimedia Companion to Anthology of Modern American Poetry.* Urbana-Champaign: University of Illinois, Department of English. Available at: http://www.english.uiuc.edu/maps/poets/a_f/frost/wall.htm.

OCLC. 2002. *A Metadata Framework to Support the Preservation of Digital Objects: A Report by the OCLC/RLG Working Group on Preservation Metadata.* (June). Dublin, Ohio: Online Library Computer Center, Inc. Available at: http://www.oclc.org/research/pmwg/pm_framework.pdf.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge: Cambridge University Press.

Ostrom, Elinor, et al. 1999. Revisiting the commons: local lessons, global challenges, *Science* 284 (April 9): 278-282.

Pauly, Mark V. 1968. The Economics of Moral Hazard: Comment. *American Economic Review* 58 (June): 531-537.

RLG. 2002. *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report.* Mountain View, Calif.: Research Libraries Group. Available at: http://www.rlg.org/longterm/repositories.pdf.

Rothenberg, Jeff. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Washington, D.C.: Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/rothenberg/contents.html.

Tenner, Edward. 2002. Taking Bytes from Oblivion. *U.S. News & World Report* 132 (April 1): 66-67.

Vaidhyanathan, Siva. 2001. *Copyrights and Copywrongs: The Rise of Intellectual Property and How it Threatens Creativity.* New York: New York University Press.

Van Orsdel, Lee and Kathleen Born. 2002. Doing the Digital Flip: Periodical Price Survey 2002. *Library Journal* (April 15): 51-56.

Waters, D., and J. Garrett, eds. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Washington, D.C. and Mountain View, Calif.: The Commission on Preservation and Access and the Research Libraries Group. Also available at: http://www.rlg.org/ArchTF/.

103

COUNCIL ON LIBRARY AND INFORMATION RESOURCES

1755 MASSACHUSETTS AVENUE, NW, SUITE 500, WASHINGTON, DC 20036-2124
Telephone: 202.939.4750 • Fax: 202.939.4765 • E-mail: info@clir.org • Web: www.clir.org

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").